Evaluation of Comparative Metabolic Network Reconstruction

Jian Hou

Helsinki, February 20, 2014
M.Sc. Thesis
Master's Degree Programme in Bioinformatics
UNIVERSITY OF HELSINKI
Department of Computer Science

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Computer Science

Tekijä — Författare — Author
Jian Hou

Työn nimi — Arbetets titel — Title
Evaluation of Comparative Metabolic Network Reconstruction

Oppiaine — Läroämne — Subject
Computer Science

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
M.Sc. Thesis	February 20, 2014	49 pages

Tiivistelmä — Referat — Abstract

Pichia pastoris and *Saccharomyces cerevisiae* are two important fungi in both research and industrial applications of protein production and genetic engineering due to the inherent ability. For example, S.cerevisiae can produce important proteins from wide ranged sugar from lignocellulose to methanol. Accurate genome-scale metabolic networks (GMNs) of the two fungi can improve biotechnological production efficiency, drug discovery and cancer research. Comparison of metabolic networks between fungi brings a new way to study the evolutionary relationship between them.

There are two basic steps for modeling metabolic networks. The first step is to construct a draft model from existing model or softwares such as the pathway tool software and InterProScan. The second step is model simulation in order to construct a gapless metabolic network. There are two main methods for genome-wide metabolic network reconstruction: constraint-based methods and graph-theoretical pathway finding methods. Constraints-based methods used linear equations to simulate the growth under your model with different constraints. Graph-theoretical pathway finding methods use graphic approach to construct the gapless model so that each metabolite can be acquired from either nutritions or the products of other gapless reactions.

In my thesis, a new method designed by Pitkänen [PJH+14] is used to reconstruct the metabolic networks of Pichia pastoris and Saccharomyces cerevisiae. Five experiments were developed to evaluate the accuracy of the CoReCo method. The first experiment was to analyze the quality of the GMNs of Pichia pastoris and Saccharomyces cerevisiae by comparing with the existing model. The second and third experiments tested the stability of CoReCo constructed under random mutation and random deletion of the protein sequence simulating noisy input data. The next two experiments were done by considering different number of phylogenetic neighbors in the phylogenetic tree. The last experiment tested the effect of the two main parameters (acceptance and rejection thresholds) when CoReCo filled the reaction gaps in the final step.

ACM Computing Classification System (CCS):
A. General Literature,
  A.1 Introductory and Survey
I. Computing Methodologies,
  I.6 Simulation and Modeling,
    I.6.3 Applications
    I.6.4 Model Validation and Analysis
    I.6.5 Model Development
    I.6.6 Simulation Output Analysis
J. Computer Applications,
  J.3 Life and Medical Sciences

Avainsanat — Nyckelord — Keywords
Metabolic Networks, CoReCo Algorithm, Evaluation

Säilytyspaikka — Förvaringsställe — Where deposited
Kumpula Science Library, serial number C-

Muita tietoja — övriga uppgifter — Additional information

Acknowledgement

At the ending point of my master study, I wish to thank many peoples for their magnanimous help. Professor Juho Rousu, my supervisor, offered me a great help for both experimental design and rectifying my thesis. Thank you for Professor Liisa Holm helped to set up GTG database. I wish to thank Dr.Merja Oja and Dr.Mikko Arvas to help me to explain and rectify my experiments. Dr.Jana Kludas, my instructor, gave me so much advices for how to evaluate the algorithm from different aspects and also helped me a lot to rectify my thesis. Dr.Esa Pitkänen, the author of the CoReCo algorithm, helped me to set up the pipeline of the CoReCo method. Kari K. Pitkänen, professor in linguistic center, gave me great suggestion on how to rectify my thesis and prepared an exercise to present my thesis. Hong Yusu, a PhD student in KEPACO group, helped me to resolve technical problems. I also want to say thank you to Professor Veli Mäkinen for his comments of my thesis.

Helsinki, February 20, 2014

Jian Hou

Contents

1 Introduction **1**

2 Metabolic Network Background **3**

2.1 Genome-wide metabolic networks 3

2.2 Model Organisms in Protein Production 5

2.2.1 Saccharomyces cerevisiae 5

2.2.2 Pichia Pastoris 6

3 Materials and Methods **7**

3.1 Drafting a reconstruction 8

3.1.1 Resources 9

3.1.2 Tools for Drafting a reconstruction 13

3.1.2.1 Constructions from existing model 13

3.1.2.2 The Pathway Tool Software 13

3.1.2.3 InterProScan 17

3.2 Metabolic network reconstruction methods 17

3.2.1 Constraint-based methods 17

3.2.1.1 Flux Balance Analysis 17

3.2.1.2 Minimization of Metabolic Adjustment 19

3.2.1.3 Flux Variability Analysis 20

3.2.1.4 RAVEN Toolbox 21

3.2.2 Graph-theoretical path finding methods 22

3.2.2.1 The Minimum mutation algorithm 22

3.2.2.2 The CoReCo algorithm 24

4 Analysis of CoReCo **30**

4.1 Original data 30

4.2 Reconstruction accuracy of P.pastoris and S.cerevisiae by ROC curves 32

4.3 Reconstruction accuracy with random sequence mutation and deletion 35

4.4 Reconstruction accuracy with varying sizes of the phylogenetic tree . 37

4.5 Reconstruction results with different acceptance and rejection parameters 39

5 Discussion **41**

References **44**

1 Introduction

The metabolic network is vital in many fields ranging from engineering projects, genetic research, drug discovery and evolutionary studies. Metabolic network is a collection of metabolic pathways. In each pathway, a series of chemical reactions catalyzed by a specific enzyme work together to accomplish biological functions: Citrate cycle pathway is an example. *Metabolites* are the chemicals in each reaction ranging from nutrition (*eg.* alcohol, sugar, lipids and amino acids) to other cofactors (*eg.* vitamin and minerals). *Genome-wide metabolic network reconstruction* is to construct metabolic network by using computational reconstruction method based on the existing evidence and annotated enzymes corresponding to the genomes of the species. Some metabolites such as antibodies are beneficial for homo sapiens. To maximize the productivity of the cell, it is necessary to know the metabolic network of the species used in production.

S.cerevisiae is the first eukaryote to be sequenced genome-wide and it was used in food production for a long time. For example, since S.cerevisiae can produce a lot of carbon dioxide during growth, it was widely used in bakeries. In research areas, S.cerevisiae is also important because it can produce many macromolecules such as different proteins, lipids and vitamins. For example, succinic acid is an essential industrial commodity because the metabolite is necessary in many chemical synthesis. S.cerevisiae can survive on a wide range of carbon resources. Production of Succinic acid of S.cerevisiae can be achieved under cheap carbon resource as nutrition to lower cost. Moreover, the inherent ability of S.cerevisiae guaranteed it can produce protein with high productivity compared with other model organisms such as Anaerobiospirillum succiniciproducens and Actinobacillus succinogenes [And92]. P.pastoris is mainly used in protein production especially to efficiently produce heterologous proteins. Moreover, it can survive on a wide range and inexpensive medium. For example, human glycosylated proteins is an important additions in pharmaceutical therapy. Due to the genetic advantage of P.pastoris (eg. strongly regulated *transcription factor binding site* (TFBS)[1], advanced post-translational regulation and the accumulated knowledge of the metabolic network, human glycosylated proteins can be safely produced by P.pastoris with high efficiency [DH05].

Complete and accurate metabolic network of S.cerevisiae and P.pastoris is necessary for both industrial production and research. Metabolic network, consisting of metabolites and enzymes, can help scientists to describe the post-translational regulation, which is not only improving the production efficiency but helping us to understand the mechanism of gene regulation and the occurrence of diseases. There are two important questions in industrial production: can we get our final production given specific nutrition? If we can, how many products we can get from the amount of nutrition? These questions can be solved only if we acquire the complete

[1]The loci located in the upstream open reading frame (ORF) of genes initiates the gene translational process included TATA box, GC box and CAAT box.

and accurate metabolic networks. In scientific research, gene knock-out technology was used to silence target gene to see whether the gene was related to the disorder. Accurate metabolic networks of the species can help us understand how does the gene affect the expression of metabolism. Metabolism is a term describing all life-sustaining reactions included enzymes and metabolites within the cells of living organisms, which maintain growth and all other biological functions.

Due to the development of innovative sequencing technology, the gap between accumulation of genomes and the metabolic networks of species is increasing. Currently, various methods have been developed to construct the genome-scale metabolic network. There are mainly two categories: constraint-based methods and graphical-based methods. Constraint-based methods use linear equations to mimic cell growth and predict functional important building blocks based on species-specific biomass Biomass is a term to describe the total amount of biological material derived from one organism in a specific period of time [PRU10]. Constraint-based methods can be used for quantitative analysis of cell growth and predict the growth rate and essential reactions. However, the speed of this method is slow, which hinders the wide usage of the methods. Graphical-based methods use a graphical approach to construct gapless metabolic network. Graphical-based approach can construct metabolic networks with lower cost (eg. less manual curation is needed). However, the accuracy is constrained by poorly sequenced data, distant homology, incorrect annotations in biological databases and missing reaction stoichiometry. It is urgent to seek a method to construct metabolic network with high accuracy and lower cost.

We use the new method, *Comparative ReConstruction* (CoReCo), to construct the metabolic network of P.pastoris and S.cerevisiae. The CoReCo is a graphical-based approach and only a little manual curation is needed to construct *gapless metabolic networks*. Gapless metabolic network means all metabolites included in the network can be acquired either from nutrition directly or from products of other gapless reactions. The characteristic of gapless metabolic network is important to guarantee we can acquire our final products based on the resource during protein production. This method takes BLAST [AGM$^+$90], GTG [HMWH07] and results from InterProScan as input and predicts the gapless metabolic networks by using Bayes network [Pea88] and atom mapping algorithm [HLMR11].

In my thesis, five experiments were developed to analyze the quality of reconstructed metabolic networks for both P.pastoris and S.cerevisiae. In the first experiment, I used ROC curves and contingency tables to test the accuracy of the reconstructed models compared to existing metabolic models that can be found in literature. The second and third experiments tested the stability of the CoReCo. For this purpose, the metabolic networks of P.pastoris and S.cerevisiae were constructed under different percentage of random mutation and random deletion of genes. The following experiment was done by considering different numbers of neighbors in the phylogenetic tree (phylogenetic neighbors). The last experiment tested the effect of the two

parameters (acceptance and rejection thresholds) in the final step of the CoReCo when using gap filling algorithm to fill the reaction gaps.

In general, the constructed metabolic networks of both S.cerevisiae and P.pastoris under each experiment show good results. The CoReCo has the ability to construct metabolic networks under poorly sequenced data. However, as incorporating more distant relatives in the phylogenetic tree (phylogenetic neighbors), the accuracy of constructed models for both S.cerevisiae and P.pastoris does not increase. The two parameters, acceptance and rejection in atom mapping algorithm, are not sensitive and can be selected in a wide range to construct models with high accuracy. A little manual curation is needed to decide whether to incorporate the reactions with big cost. In Chapter 2, I will introduce some definitions and two well-studied fungi: P.pastoris and S.cerevisiae. In Chapter 3, I will introduce the state-of-art methods to construct metabolic networks and the CoReCo method used in my thesis. In Chapter 4, I will present the data in my thesis and show the results of the experiments. Finally, I will discuss the result in Chapter 5.

2 Metabolic Network Background

2.1 Genome-wide metabolic networks

Genome-wide metabolic network is a collection of metabolic pathways. Each metabolic pathway consists of chemical reactions and enzymes catalyzed for specific reactions to accomplish functions: TCA cycle is an example (Figure 1). Different metabolic pathways work together to sustain the normal physiological and biochemical properties of a cell. For example, Acetyl-CoA is generated from four other metabolic pathways and served as substrate in TCA circle at first. Then, a set of chemical reactions joins together based on Acetyl-CoA to produce the NADH (Nicotinamide adenine dinucleotide). Finally, NADH is produced and transferred to mitochondria and served as substrate in the aerobic respiration pathway to produce energy. Many databases of metabolic pathway have been established such as BIOCYC, MetaCYC [CAD12], ExPASy and KEGG [KG00], which offer us a great opportunity to accurately predict metabolic network in genome-scale.

Genome-wide metabolic network is important because it is not only describing function of the proteins but also help us to understand the protein-protein interactions in metabolism, which helps us to study the occurrence and development of diseases [PN05]. Metabolic network has been used in other fields except the study of protein-protein interactions. For example, accurate-metabolic network can help us to annotate genes. Based on sequence homology and the annotations of genes of species from previous studies and databases, we can predict metabolic network of unknown species [Ost03]. Moreover, production efficiency of microorganisms is

we can predict for metabolic netwo

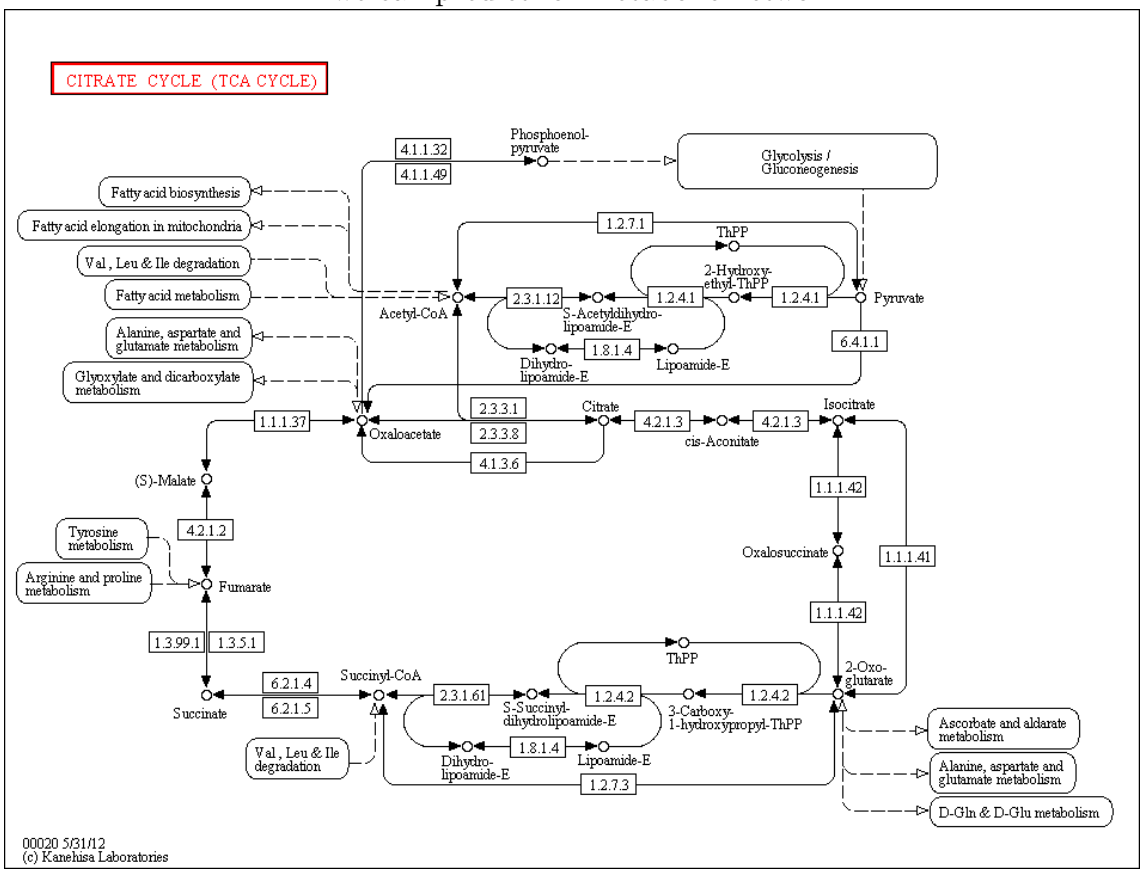


Figure 1: The Citrate cycle (TCA cycle) reference pathway from KEGG database. Solid arrows and lines are the chemical reactions and directions within the TCA pathway. Dot lines and blank arrows are the reactions and directions outside the TCA pathway. Small circles are the metabolites and each four digit number within rectangle express one unique enzyme. Products from other pathways (Fatty acids metabolism for example) serve as input metabolites in the TCA circle and vice versa.

quite essential in protein production. Comparison of metabolic networks to identify the functional differences can offer us a better choice when choosing microorganism to maximize productivity.

2.2 Model Organisms in Protein Production

Model organism are non-human species that have been extensively studied to understand particular biological phenomenon and enable to provide insight discovered in the model organism to other species. Both S.cerevisiae and P.pastoris have been well-studied in research such as human disease, drug discovery and evolution. Many results were found by study of S.cerevisiae. For example, several genes related with aging have been identified by study of S.cerevisiae [RBCTR04]. Recently, due to the genetic advantage, P.pastoris was frequently used in *heterologous protein*[2] production.

2.2.1 Saccharomyces cerevisiae

S.cerevisiae has been widely used in food processing related to bakeries and wine-making. For example, the oldest beverage fermentation using S.cerevisiae was found round 7000 BC in ancient China, 3000 BC in Egypt and 6000 BC in Iran [LMCK07]. In addition to food production, S.cerevisiae is also popular in commercial applications to produce lipids, proteins and vitamins. In 1986, S.cerevisiae was first considered as harmless organism by Food and Drug Administration and Department of Health and Human Services (DHHS). Moreover, after several authorities published S.cerevisiae as a safe organism, it had been frequently used in protein production. Currently, as one of the most important eukaryote, S.cerevisiae is nominated as the primary model for genetic studies.

S.cerevisiae is one of the most studied fungi with a long history. S.cerevisiae was commonly used in protein production because of its unique characteristics. Traditionally, the majority of protein production and purification have been made by using prokaryote such as E.coli. However, to produce heterologous proteins from advanced species successfully, complicated post-translated modification is compulsory in order to form functional proteins. Meanwhile, the expression system should not be too complicated to manipulate. S.cerevisiae satisfied these conditions. S.cerevisiae is a eukaryote. The advanced post-translational modification system guarantees the heterologous protein can be produced with S.cerevisae. Moreover, compared with other model organism, fewer toxins and other self-proteins were produced by S.cerevisiae during protein production, which makes it easy for protein purification.

[2]Heterologous proteins are the proteins that translated from the gene that belongs to other species. These genes were first constructed into vectors (e.g plasmid) and following integrated into the target genome (e.g P.pastoris) by genome reconstruction technology [RSC92].

Finally, the existing knowledge of metabolic network of S.cerevisiae supports and guarantees the wide usage of S.cerevisiae in both commercial protein production and genetic research.

The study result of S.cerevisiae helps us to identify gene functions for unknown species. The functions of many genes of homo sapiens were primarily deduced from study of S.cerevisiae. For example, three mutated genes of S.cerevisiae (PMS1, MLH1 and MSH2) have been proved to cause trait instability (fragile chromosome that break easily) on S.cerevisiae chromosome. Moreover, similar genes were found in homo sapiens by sequence homology between S.cerevisiae and homo sapiens, which suggested the potential mutations may be associate with colorectal cancer [SPLP93]. Werner's syndrome is a disease that has common performances with *premature aging*. Premature aging is a phenomenon often observed in a set of rare hereditary (genetic) disorders. The symptoms often related to accelerated aging: skin wrinkle is example, which is often related with DNA damage. Depending on the existing knowledge, SGS1 gene of S.cerevisiae is closely regulated with its life span. Based on sequence similarity, Sinclair *et al* [SMG97] successfully discovered the homologous gene of SGS1 in patients. Further experiments identified the SGS1 gene location and function in the metabolic pathway, which suggested the potential gene of Werner's syndrome [SMG97]. Finally, there were some examples suggested S.cerevisiae is commonly used in commercial production such as enzymes, antibodies, lipids and vitamin (Table 1).

Table 1: Heterologous proteins expressed in S.cerevisiae

Protein	Species	Reference
virus polyprotein (VR2 and VR3)	virus	[NVI$^+$90]
Newcastle disease virus Matrix proteins (NDV-M-protein)	virus	[ISI$^+$13]
lipase (LIP2)	fungi	[Dar12]
CgAqr1	fungi	[CHP$^+$13]
isoprene (IspS)	plant	[HZM12]
Aquaporin-1	homo sapiens	[BHNSPP13]
N-glycosylation recombinant glycoproteins	homo sapiens	[ABJ13]
Hepatitis B core protein (HBcAg)	homo sapiens	[BFC$^+$90]
asparagine synthetase	homo sapiens	[vHS90]
Human-IFN-α	homo sapiens	[HHL$^+$81]

2.2.2 Pichia Pastoris

P.pastoris was known as one of the most important yeast in heterologous protein production for a long time. In 1970s, Koichi Ogata first used P.pastoris to produce single cell protein (SCP) for animal feed. In 1980s, Phillips Petroleum who worked in the SIBIA company first developed the heterologous gene expression system of P.pastoris, which could incorporate those foreign genes with high commercial price

into P.pastoris genome and expressed the genes with high productivity.

Several factors contributed to the fact that P.pastoris is widely used in heterologous protein production [CC00]. Firstly, P.pastoris is a single cell microorganism that is easy to manipulate compared with multi-cell organism. Moreover, similar to S.cerevisiae, P.pastoris has advanced *post-translational modifications*. Post-translational modifications are some chemical reactions related to glycosylation, folding and proteolytic processing, which is necessary to form functional proteins. This characteristics guarantee proteins from eukaryotes can be produced in P.pastoris. The original proteins are non-functional when it translated from mRNA. Heterologous protein were functional only after processing proper post-translational modifications: to conform the correct 3D structure for example. Thirdly, methanol is consumed as the major carbon resource of P.pastoris to offer energy, which decreases the price for protein production. Fourthly, the TFBS of P.pastoris in the promoter region of *alcohol oxidase 1* (AOX1) gene is highly related with methanol metabolism, which sufficiently improved productivity during protein production. Fifthly, the similarity between S.cerevisiae and P.pastoris makes it easier to transfer the same technique from one species to another. Finally, the inherent ability of P.pastoris that strongly preferred to respiratory growth facilitated to cultivate P.pastoris with high-cell density relative to other yeast.

Compared with S.cerevisiae, there are three main reasons contributing to choose P.pastoris for heterologous protein production [DS96]. Firstly, no strongly inducible promoters were found in S.cerevisiae. Methanol is the major carbon resource to offer energy for P.pastoris but not for S.cerevisiae. The TFBS located in the AOX1 gene is restrictively regulated during methanol metabolism, which increases the heterologous protein productivity. Thirdly, post-translational modifications of P.pastoris is more similar with homo sapiens than S.cerevisiae meaning most of human proteins from P.pastoris is functional but not in S.cerevisiae. Finally, less toxic and self-protein were produced during heterologous protein production by P.pastoris, which make easy for purification. Many successful cases are published for heterologous protein production by P.pastoris such as Mature sakacin A, Staphylococcal Protein A and xylanase (Table 2).

3 Materials and Methods

,

Table 2: Heterologous proteins expressed in P. pastoris

Protein	Species	Reference
Mature sakacin A (SakA)	bacteria	[JBD+13]
Staphylococcal Protein A (SPA)	bacteria	[HXH+13]
xylanase (xynB)	fungi	[FGCS13]
Aspartic Protease	fungi	[YCSZ13]
Ganoderma Lucidum TR6	fungi	[YLL+13]
Amino Acid Transporter ASCT2 (hASCT2)	homo sapiens	[PPS+13]
Human Serum Albumin (HSA)	homo sapiens	[WWL+13]
EBNA1	virus	[WJLW13]
Equine Infectious Anemia Virus (EIAV) Antigen	virus	[CJF+13]

3.1 Drafting a reconstruction

The first step of metabolic network modeling is to construct the draft model. The quantity of the biological annotations for the specific organism and the biochemical database directly determines the model quality [TP10]. Currently, due to the development of sequencing technology [Met09] and accumulation of existing knowledge [KG00, KOMK+05, Bai00], there are a quite few of resources available to semi-automatic construct metabolic network. *Comparative genomic approach* was used to construct the draft metabolic network: protein sequences search against a reference database by *sequence homology*[3] is an example. Draft network often have some "gaps" due to missing enzyme that are necessary for the catalyzation or metabolites that cannot be reached neither from resource nor other "gapless" reactions. Gaps need to be filled for several reasons [OP12]. Gaps block the reactions that makes some metabolites non consumable or producible, which influence the result of model simulation. Moreover, the process of filling gaps may help us identify new genes and functions of metabolites. Gaps in the model make it difficult for quality control during protein production of industry. Gaps have to be filled to guarantee products can be acquired based on the predicted metabolic network. Several methods were used to fulfill this goal in order to fill these gaps: find functional coupling is an example. Functional coupling are the proteins that seem interacted together and the reactions should exist in the models. Finding functional coupling offers a way to extend our reactions pools that may fill reactions gaps in the draft model.

[3]Similarity between sequences. There are two main kinds of homology: orthologs and paralogs. Orthologs are for the sequence that directly transferred from the ancestor to the posterity. Orthologs genes have similar functions in different species. Paralogs are for the sequence that acquired from gene duplication. The duplicated genes have similar sequence and chemical structure. However, these duplicated genes often do not have similar functions between the ancestor and the posterity: missing transcription factor binding sites is an example that makes the paralogs genes fail to transcript from DNA to mRNA. In this work, we don't distinguish orthologs and paralogs. Instead, we assume sequences with high similarity have similar functions but it is not necessarily true.

3.1.1 Resources

There are a lot of resources that offer a great opportunity for the semi-automatic assembly of the metabolic network. Kyoto Encyclopedia of Genes and Genomes (KEGG) [KG00] is an integrated database that contains information of genes, proteins, reactions and pathways. KEGG is prominent in understanding of high-level functions such as protein interaction and study metabolic pathways. For example, the most unique data in KEGG is the metabolic networks, which describe molecular interaction, reaction and related networks extracted from literature and recorded into three different databases: KEGG PATHWAY, KEGG BRITE and KEGG MODULE. In KEGG PATHWAY, each pathway is species-specific categorized and various data objects are available such as genes, proteins, reactions, metabolites and reported literature. KEGG database was first built in 1995 by Kanehisa Laboratories and consists now of sixteen main databases (Table 3).

Table 3: Main databases in KEGG

Category	Database	Content
Systems information	KEGG PATHWAY	KEGG pathway maps
	KEGG BRITE	BRITE functional hierarchies
	KEGG MODULE	KEGG modules of functional units
	KEGG DISEASE	Human diseases
	KEGG DRUG	Drugs
	KEGG ENVIRON	Crude drugs and health-related substances
Genomic information	KEGG ORTHOLOGY	KEGG Orthology (KO) groups
	KEGG GENES	Gene catalogs in complete genomes
	KEGG SSDB	Sequence similarity database for GENES
Chemical information	KEGG COMPOUND	Metabolites and other small molecules
	KEGG GLYCAN	Glycans
	KEGG REACTION	Biochemical reactions
	KEGG RPAIR	Reactant pair chemical transformations
	KEGG RCLASS	Reaction class defined by RPAIR
	KEGG ENZYME	Enzyme nomenclature

BioCyc is another powerful database that integrated 2988 pathway/Genome Databases (PGDBs). Each PGDB collects information of genomes and metabolic pathway for a single organism. BioCyc consists of six intensively curated databases (Table 4). For example, MetaCyc is a group of metabolic pathways and enzymes for multi-species containing 2042 metabolic pathways from 2414 species extracted from 36329 publications. MetaCyc provides reference data such as pathways, reactions, enzymes and metabolites to support metabolic network reconstruction and serve as an encyclopedia of metabolism [CAD12]. BioCyc also contains many tools for visualizing and analysis of *omics*[4] data: Genome browser for example. Moreover, the PATHWAY

[4]A term refers to the biological data ending with omics, such as genomics, proteomics and metabolomics. For example, genomics means the whole DNA sequence for one species.

TOOL in BioCyc offers a convenient way to semi-automatic create your own PGDB and build a *flux balance model* to evaluate productivity of your PGDB.

ENZYME is a repository database of multi-species enzyme [Bai00]. Each pro-

Table 4: Main databases in BioCyc

Database	Scope	Content
EcoCyc	E.coli metabolic pathways	Complete genome Transcriptional regulatory network Flux balance model
MetaCyc	Multiorganism Metabolic pathway and Enzyme database	2042 metabolic pathways from 2414 organisms
HumanCyc	Homo sapiens	297 metabolic pathways
AraCyc	Arabidopsis thaliana	400 metabolic pathways
YeastCyc	S.cerevisiae	152 metabolic pathways
LeishCyc	Leishmania major Friedlin	143 metabolic pathwaysÂ

tein is categorized by its chemical property and describes with *Enzyme Commission* (EC) number. The EC number is a numerical classification for each enzyme. Each full EC number describes by four digit number. Because EC number classifies enzymes by their categorized reaction instead of a specific enzyme, different enzyme may have the same EC number. ENZYME database is convenient especially for filling reaction gaps in the draft metabolic network. For example, some proteins are necessary to fulfill a biological function but do not predict in the draft model by sequence homology. ENZYME can be used to query these proteins to see whether there is any evidence to support the existence.

Swiss-Prot is a central hub for the collection of functional informations on genes, proteins, classification and cross-reference of publications [BBA$^+$03]. Moreover, Swiss-Prot is highly nominated as the gold standard for computational system biology due to the fact that a well-defined manual curation is processed to guarantee the accuracy and consistency of annotations. The main process of manual curation is divided into three steps. First, new sequence is blast search against the Swiss-Prot database to find the discrepancy annotations between the new report and record. Many reasons cause the discrepancy such as mutations, *alternative splicing*, incorrect exon boundaries and initiation sites. Alternative splicing is a process happening on DNA transcription. During this process, a modification of pre-mRNA transcripts in which introns are removed and exons are combined in different manners. Alternative splicing can produce a range of proteins from the same gene by varying the composition of exons. The next step is literature curation by identification of other experimental evidence such as gene name, function, cofactors, subcelluar location, protein protein interactions and other biological and chemical related information to guarantee the non-redundancy and consistency of the sequence. Finally, *recip-*

rocal blast search[5] and other phylogenetic resources are used to determine putative homologous region.

The global trace graph (GTG) database by Heger [HMWH07] is created to find the homologous proteins. In particular, GTG database is more sensitive to detect the homologous sequence of distantly related species. The original sequences were chosen from different databases such as Swiss-Prot, PDB, Protein Data Bank, SCOP and PFAM. ADDA algorithm [HH03] were designed to find the homologous domain and *nrdb40* was created after excluding the redundant sequences with more than 40% similarity detected by ADDA algorithm. Next, the unweighted alignment trace graph was created based on nrdb40 by multi-sequence alignment. In the final step, the alignment trace graph was weighted by the consistency of residues between two neighbors in the graph node. Given a set of protein sequences, searching against GTG database will return seven features (Figure 2). *QID* is the identification of

QID	QP	QA	MID	MP	MT	MF
Q0U8Z5	1	M	66969	1	11	74561101
Q0U8Z5	2	S	66969	2	16	46139615
Q0U8Z5	3	T	66969	3	17	41261867
Q0U8Z5	5	K	66969	5	9	58558121
Q0U8Z5	6	I	66969	6	8	29985620
Q0U8Z5	7	T	66969	7	17	59650952
Q0U8Z5	9	L	66969	9	10	27280920
Q0U8Z5	10	T	66969	10	17	17722529
Q0U8Z5	11	N	66969	11	12	37255944
Q0U8Z5	12	W	66969	12	19	38549513
Q0U8Z5	14	A	66969	14	1	62104137
Q0U8Z5	15	T	66969	15	17	30470412

Figure 2: The original file from searching against GTG database. There are seven features in this picture. The total number of conserved features (marked by "MF") of each matched sequence (marked by "MID") can be used to evaluate the similarity between each query sequence with the matched sequence.

the query sequence, *QP* is the position of the query sequence, *QA* is the amino acid of the query sequence. Similarly, *MID, MP, MA* is the identification, position and amino acid of the matched sequence. *MF* in the last column was the matched feature. The total number of the matched feature was used to evaluate the similarity between each query and the matched sequence.

Database Expressed Sequence Tag (dbEST) [BLT93] is a division of Genbank established in 1992. As for GenBank, data in dbEST is directly submitted by lab-

[5]Reciprocal blast search is a two-directional blast searching process. Query sequences first search against a database and then sequence of database will search against the query. Matches are only recorded if targets are found in both blast results.

oratories worldwide and is not curated. Some experiments use EST[6] to evaluate their metabolic model and further increase accuracy. For example, Wanwipa *et al* [VOH+08] reconstructed the model of A.oryzae from the extended gene pool by considering both the genes from existing A.oryzae database and an EST library of A.oryzae. The high quality EST reconstruction result describes the transcriptome of A.oryzae and was utilized for gene prediction by mapping the assembling ESTs to the A.oryzae genome (Figure 3). Furthermore, with the extended gene pool, candidate proteins and reactions of each gene of A.oryzae was extracted from the result of sequence homology among A.nidulans, A.fumigatus and S.cerevisiae. Finally, an integrated tool (GFAOP) was designed to infer the possible reactions to fill the gaps in predicted metabolic network of A.oryzae. The method combined different genetic informations ranging from the genome, transcriptome and functional annotations of each protein that improved the prediction accuracy. However, extensive manual curation of the gap filling for the final reconstruction step and strong dependency of annotations of S.cerevisiae, Anidulans and A.fumigatus hindered the wide usage of the method. In conclusion, the scope of different databases is shown in Table 5.

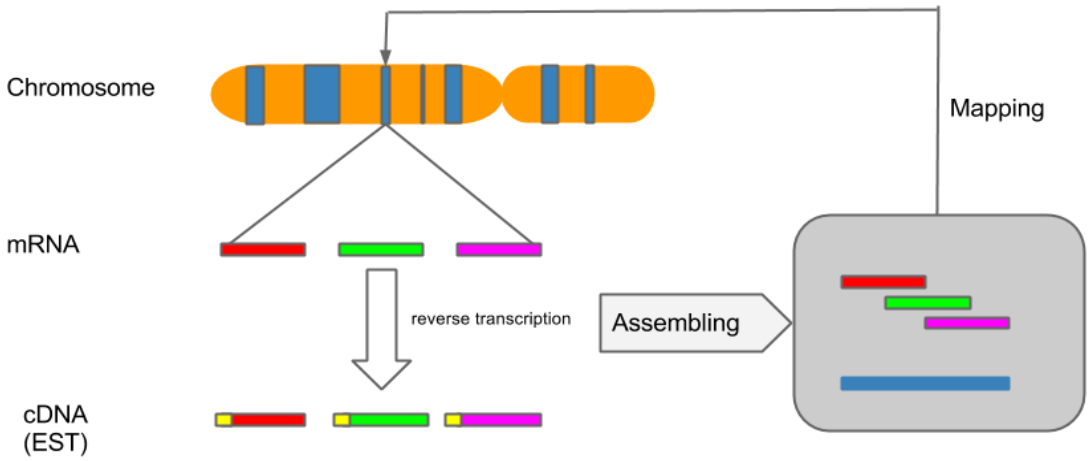


Figure 3: Diagram descriptions of assembling EST and mapping to the genome. Blue color on chromosome expression genes located on the chromosome. Several mRNAs are first transcript from the same gene. Each mRNA are reverse transcript into complementary DNA (cDNA) and add primer sequence at one end (yellow), which is called EST. Finally, all ESTs are assembled by reconstruction software (eg. *Cisgenome*) and mapped to the location of the reference genome.

[6]EST are short cDNAs that reverse translation from mRNA to DNA and exclude untranslated regions compared with ORFs on the genome. Hight quality of EST described gene expression level can be used to test and improve your model.

Table 5: Comparison of different databases

Database	Enzymes	Genes	Reactions	Pathways	Metabolites
KEGG	YES	YES	YES	YES	YES
BioCyc	YES	YES	YES	YES	YES
MetaCyc	YES		YES	YES	YES
ENZYME	YES		YES		YES
Swiss-Prot	YES	YES			
dbEST	YES	YES			
GTG	YES	YES			

3.1.2 Tools for Drafting a reconstruction

3.1.2.1 Constructions from existing model

Organism-specific database serves as a critical resource to acquire metabolic data (e.g. reactions, enzyme, conditions of experimental data and gene expression). This is extremely useful when evaluating the quality of your result by comparing the predicted growth rate with the experimental results under certain conditions. Alternatively, if organism-specific database is not available, genomic annotations can be extracted from their relatives by sequence homology and Gene Ontology [ash]. Close relatives with similar genomes and/or proteomes may have similar biological functions. In case of Luis Caspeta *et al* [CSA+12] metabolic network construction of P.pastoris, the new proteome sequence was searched against the existing P.pastoris models (GS115). P.stipitis as one of the closest relatives of P.pastoris was also incorporated to extend the gene pools and biological functions that might be missed due to missing, wrong or incomplete annotations. Another example of drafting metabolic networks is by Helga David *et al* [DÖHN08]. In this example, annotations of A.nidulans[7] were not only extracted from the relatives. To incorporate more biochemical informations, sequence homology were also processed by comparing A.nidulans with evolutionary distant species such as mouse, rat and homo sapiens(Figure 4).

3.1.2.2 The Pathway Tool Software

The pathway way tool software is powerful to quickly build a *Pathway/Genome Database (PGDB)* and further drafting metabolic models with less effort [KPR02]. Moreover, the pathway tool provides a convenient way for scientists to interrogate the PGDB with complicated querying and visualize the query in an intuitive, graphical fashion. There are mainly three components of the pathway tool: the pathologic pathway predictor supports to create a new PGDB from the annotated genome of an

[7]One type of fungi which has well-known as an important research organism for studying eukaryotic cell biology.

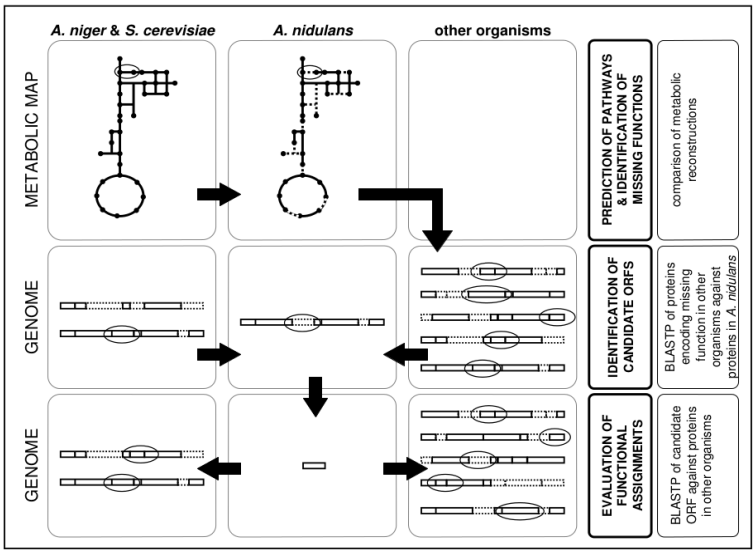


Figure 4: Diagram descriptions of the reconstruction process of A.nidulans. Annotations of A.nidulans were first incorporated based on the annotations of S.cerevisiae by sequence homology. Distant species (middle right) were then also incorporated into comparison to extract more biological informations that were not identified in the first step. Those new identified annotations were evaluated and assigned to the A.nidulans draft model.

organism. The Pathway/Genome Navigator provides complicated query and visualization service of PGDB. For example, given an individual entity such as metabolites, enzymes or pathways, Pathway/Genome Navigator returns the result and highlights the results specific or all pathways. Moreover, user can visualize the entire pathway under different levels such as metabolites, the chemical structures of substrates and the genes associated with each enzyme (Figure 5). The Pathway/Genome Editor supports interactive updating of PGDBs. To create a PGDB of new species S.cerevisiae, a set of flat file format files is collected containing annotations such as gene location, gene name, type of the gene product, sequence and EC number. The input flat file can be various: GeneBank is an example (Figure 6). The Pathologic pathway predictor first read the set of collection and transfer to PGDB format for further updating and visualization. Moreover, the collected annotations of organism would be compared with the annotations in all pathways of the existing PGDB (see Resources) to acquire significant evidence on which pathway should be incorporated into this organism. Because the pathway tool pretends to incorporate more pathways rather than only collecting those with strong evidence, false positive pathways would be considered into the drafting model. Updating the existing PDGM such as incorporating more experimental evidence and other manual curation are necessary to increase the accuracy of the model.

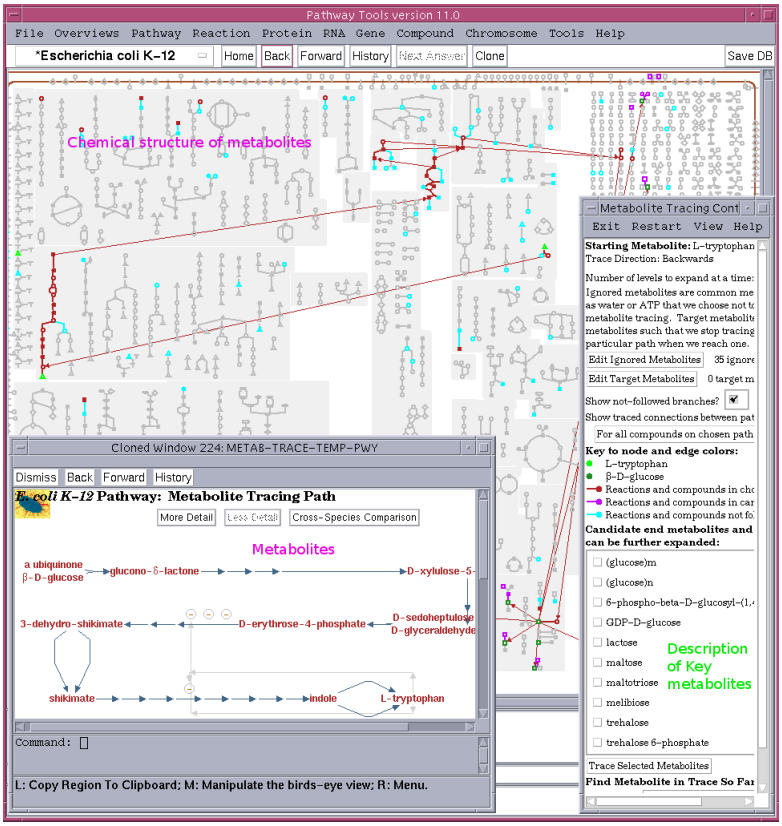
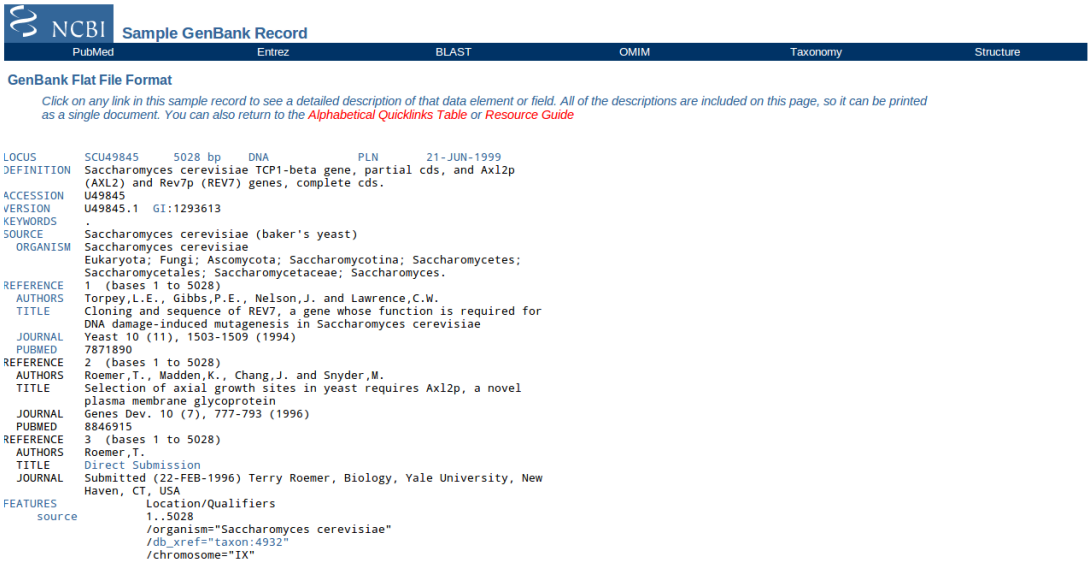


Figure 5: Visualization of E.coli K-12 pathway on different levels. The E.coli K-12 pathway is described in metabolites (bottom left), chemical structures (top). Descriptions of the key enzyme and metabolites in the E.coli K-12 pathway is colored by red and described on the right panel.


NCBI Sample GenBank Record

PubMed Entrez BLAST OMIM Taxonomy Structure

GenBank Flat File Format

Click on any link in this sample record to see a detailed description of that data element or field. All of the descriptions are included on this page, so it can be printed as a single document. You can also return to the Alphabetical Quicklinks Table or Resource Guide

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
```

Figure 6: This is a Genebank example file. *LOCUS* is the Chromosome location; *DEFINITION* is the description of the gene; *ACCESSION* is the accession number in Genebank. *ORGANIMS* is the species on which the gene that belongs to; From *REFERENCE* to *PUBMED* is the cited information of the gene; *FEATURES* contains the description of the gene; *source* contains the number of nuclear acids of the gene; *CDS* is the number of coding areas of the sequence; *ORIGIN* is the sequence of the gene.

3.1.2.3 InterProScan

InterProScan is a tool to predict protein domain and functions by comparing each protein sequence against 14 protein signature databases [ZA01]. There are two ways to use InterProScan. For functional prediction of several protein sequences, it is convenient to access InterProScan via a web server (Figure 7). Different kind of input such as InterProScan ID, GO term ID, protein sequence and type of protein domain can be used for searching. The result contains many informations such as family of protein domain, functional predictions, PubMed ID and GO term ID. For big data set, InterProScan has a standalone version. Users can download this version from http://ftp.cbi.pku.edu.cn/pub/software/unix/iprscan/. The standalone version can run multi-sequence in parallel and provides several popular output formats. The simple retrieval system makes it easier to transfer among different output types.

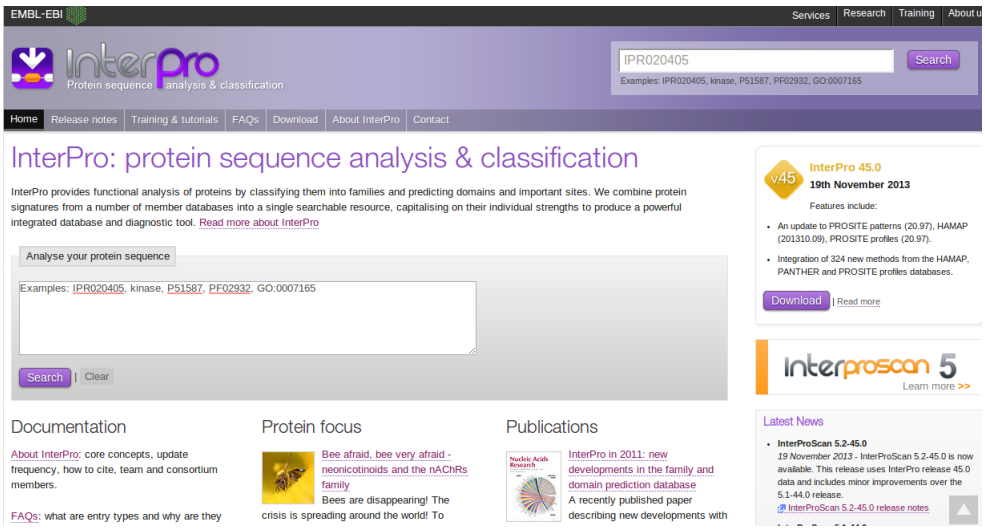


Figure 7: InterProScan web server homepage. There is different type of entry for searching such as InterProScan ID (eg. IPR020405), type of protein domain (eg. kinase), GO term ID (eg. GO:0007165) or protein sequence.

3.2 Metabolic network reconstruction methods

3.2.1 Constraint-based methods

3.2.1.1 Flux Balance Analysis

Flux Balance Analysis (FBA) is a mathematical method to simulate cell growth or differentiate. The advantages of FBA such as computationally inexpensive and easy to incorporate more constraints make it commonly used in several applications such as identifying putative drug target in cancer and pathogens, evaluation of

connectivity and growth rate of draft model, identify essential reactions and genes [HDB+10]. The formula is described in the following equations.

$$max \quad c^T V \tag{1}$$
$$s.t \quad S \cdot V = \mathbf{0} \tag{2}$$

where v_x is a specific component of V, S is the stoichiometric matrix that rows are metabolites and columns are reactions and V is the *flux*. Flux is a vector that describes the production rate of each reaction under the steady state where all resources are transferred into products (equation 2). Active components of flux are products and negative components are consumptions supplemented from environment such as carbohydrates. For example, the method designed by Borodina *et al* [BKN05] used FBA analysis to estimate the growth rate of the predicted model. The first step is to create draft model by combining genome annotation from several databases (eg. KEGG Pathway database, The Welcome Trust Sanger Institute database, KEGG Ligand database, ExPASy Biochemical Pathways, SWISS-PROT database) and evidence from literature to predict the draft model of S.coelicolor. The productivity and connectivity [8] of this model are tested and verified by flux balanced analysis and essential genes and reactions of S.coelicolor were also identified [BKN05].

A lot of suggestions are described by incorporating more constraints based on flux balance analysis. For example, Vinay Satish *et al* [KDM07] creates an approach to systematically identify gaps and fill these gaps by extending flux balance analysis. This approach is divided into two subprocesses. *GapFind*, a computational based method, is first used to detect the metabolites that could not be produced from the original metabolic model by FBA. After identification of the gaps, *GapFill* is used to fill the gaps. GapFill first considers whether the gaps could be filled after changing the directionality of the reactions in the draft model. For example, the reaction iJR904 in the metabolic model of E.coli is registered as only one direction. However, the reaction is listed as reversible in the database of EcoCyc and the metabolite can be reached after reversing the directionality of this reaction. Gaps are related with the metabolites that either cannot be consumed or cannot be generated during model simulation. Once gaps were identified, annotations of these metabolites and reactions from databases such as KEGG can help us to fill these gaps. If the gaps are not resolved after reversing the directionality, new reactions are considered to be added to this model. Reactions are acquired from the Metacyc database and incorporated into this metabolic model if it does not exist in the metabolic model. Finally, a new flux analysis is processed to fill the gaps while trying to maintain the structure of the existing metabolic network. In particular, the objective function of

[8] Productivity and connectivity are two important parameters for *in silico* growth simulation, which describe the ability of producing a set of target proteins.

the new flux analysis is to minimize the number of new reactions (equation 5).

$$minimize \quad \sum_{j' \in Database} y_{j'} \tag{3}$$

$$\begin{aligned} s.t. \quad & LB_j \leq v_j \leq UB_j \forall j \in Model & (4)\\ & LB_{j'} \cdot y_{j'} \leq v_{j'} \leq UB_{j'} \cdot y_{j'} \forall j' \in Database & (5)\\ & S \cdot V \geq \mathbf{0} & (6) \end{aligned}$$

where y_i are the new reactions that incorporate into this model, LB_j and UB_j are the lower and upper bound of the reaction existing in the metabolic model, $LB_{j'}$ and $UB_{j'}$ are the lower and upper bound of new reactions belong to the Metacyc database, $S \cdot V \geq \mathbf{0}$ is similar with constraint in Equation 2 but to minimize the growth process instead of computing the maximum growth rate. This is because the aim of the constant (Equation 6) is to guarantee the model had the ability to produce the metabolite rather than computing the maximum growth rate described by equation 2 under the assumption of all of the input resources have been transferred to the products. This method offers a possible way to identify the reaction gaps and resolves the puzzle of how to fill the gaps. In particular, this approach does not need much manual curation thus reconstructions of models of multi-species can be processed in parallel. However, the disadvantage of this method is that the reconstruction process is mandatory built on the existing models. Moreover, the ability of this method to fill the gaps is much dependent on the reactions registered in the Metacyc database.

3.2.1.2 Minimization of Metabolic Adjustment

Minimization of Metabolic Adjustment (MOMA) [SVC02] is another method to estimate the growth rate similar with flux balance analysis. MOMA have the similar constraints but different objective function. The formula of MOMA is outlined in the following equations.

$$\begin{aligned} min \quad & \|V_o - V\|^2 & (7)\\ s.t. \quad & S \cdot V = \mathbf{0} & (8) \end{aligned}$$

where V_o is the optimal flux vector and V is the mutant flux vector. The constraint of MOMA is the same as the constraint in the FBA (see Equation 2). Instead of predicting the maximum growth rate, MOMA loose the boundary to calculate a reasonable range of the mutant flux that close enough to the optimal FBA flux. This is more useful when you compare your expectation with experimental data. For example, Montagud *et al* [MNdC+10] constructs the draft model of Synechocystis by incorporating annotations from publications and several databases such as KEGG and BioCyc. To compute the growth rate and compare with experimental

data, FBA was first used to simulate the maximum productivity under different carbon resources. However, the maximum growth rate is not consistent with the experimental result in the beginning and finally the two results were matched after a continuing sub-cultivations of Synechocystis for 40 days. This suggests a function that can predict the growth in a reasonable range rather than only simulations of the optimal growth is required. In the new test, the author used both the MOMA and the FBA methods to simulate the network and compared with the experimental data. The result was consistent with most of the data from the beginning of the experiment.

3.2.1.3 Flux Variability Analysis

Flux Variability Analysis (FVA) is another method used to evaluate metabolic models. Instead of maximizing flux of all reactions in FBA, FVA tend to identify the minimum and maximum flux for a single reaction in the network while maintaining some state of the netwok, eg. supporting 90% of maximal possible biomass production rate. FVA has been used in many applications such as identifying active, inactive or essential reactions, studying flux distributions under suboptimal growth, identifying functional important building block and investigating network flexibility and network redundancy [GT10]. The formula is showed below

$$min/max \qquad v_i \tag{9}$$

$$s.t. \quad S \cdot V = \mathbf{0} \tag{10}$$

$$w^T V \geq r Z_0 \tag{11}$$

$$v_l \leq v \leq v_u \tag{12}$$

, where S, V is the same as in FBA, v_i is one component in flux vector V, w is the the matrix we want to compute, Z_0 is the optimal solution from FBA, r is the parameter ranging from 0 to 1. Specifically, $r = 1$ is an optimal solution to (1). This is the because Z_0 is the optimal flux produced from FBA. For anySeveral examples prove FVA can help us to identify essential genes, reactions and functional important blocks. For example, Henry *et al* [HDB$^+$10] developed software *SEED* that automatic construct metabolic networks. The pipelines consist of three steps (1) run sequence data on the RAST server and import the annotations into SEED. During this process, the Gene-Protein-Reaction (GPR) table had been created. Moreover, species specific cofactors had been identified based on gene annotations.(2) To fill gaps of the draft model, FBA was first used to identify the non-generate biomass reactions. Then an optimization algorithm was used to determine the minimum set of reactions that must be added into the model to fill these gaps. (3) In the final step, FVA was used to classify reactions into three categories namely active, essential and inactive reactions. Removing those inactive reactions from models increased prediction accuracy from 10% to 20%.

3.2.1.4 RAVEN Toolbox

RAVEN Toolbox by Agren *et al* [ALS+13] is an powerful tool that uses constraint based method to create metabolic models with few manual curation. In RAVEN Toolbox, *Hidden Markov Model* was used as similarity tool to fill gaps instead BLAST. The first step of RAVEN was to create the raw metabolic network based on protein homology. Two strategies of homologous searching were suggested. To maximize utilizing the metabolic informations based on the existing models, models of three fungi that closely related to P.chrysogenum were considered as temples and all reactions were added to the raw metabolic network of P.chrysogenum. To fill gaps of the draft model, the *KEGG model* was processed in following steps. These gaps were identified based on the full network of subcategory (eg. eukaryotes or prokaryotes) from KEGG. Reactions, genes and metabolites related with gaps would search in *KEGG Ontology* (KO). KEGG Ontology is an online tool that clusters genes of all species into different categories based on cellular compartments, molecular functions and biological processes. Each KO term are composed by the genes, reactions, cellular processes and human diseases (Figure 8). A set of proteins

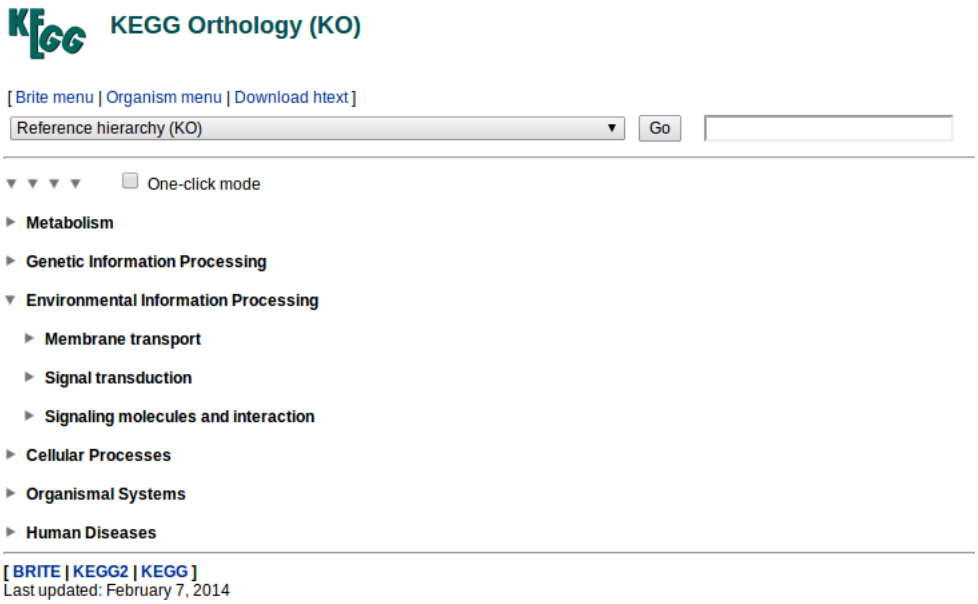


Figure 8: An example KO term of p53 protein.

would acquire and align by *multi-sequence alignment tool* (MUSCLE). Finally, a set of HMM would be created for these protein and reactions would be incorporated if sequences of P.chrysogenum had high similarity with the HMM. The result showed RAVEN Toolbox has the ability to construct metabolic models with high accuracy and less manual curation based on existing knowledge. As more metabolic models are available, the prediction accuracy may increase.

3.2.2 Graph-theoretical path finding methods

3.2.2.1 The Minimum mutation algorithm

Pitkänen *et al* [PRA11] constructed gapless metabolic networks by using graphical-based approach while minimizing the total mutation cost. The draft model of a group of species was predicted by extracting reactions, metabolites and other annotations from the KEGG database. The algorithm divides into two main processes. To create the gapless metabolic network with minimum mutations, Fitch algorithm was first used to compute the minimum mutations given the phylogenetic tree structure based on the evolutionary relationship between the species (Figure 9). Fitch algorithm first computed each reaction of internal nodes from the bottom of the phylogenetic tree to the top based on equation 13.

$$F_i(v) = \begin{cases} \{L_i(v)\} & \text{if } v \text{ is leaf} \\ F_i(x) \bigcup F_i(y) & \text{if } v \text{ is not leaf; x and y are the left and right child of v} \end{cases} \tag{13}$$

where $F_i(v)$ mean the existence of the reaction i in the internal node v; $L_i(v)$ means the existence of reaction i in leaf node v. Leaf nodes are the kind of nodes that contain only parent nodes but do not exist child nodes. If the reaction exists in the leaf node, it will be marked with 1, otherwise marked with 0. For an internal node, the existence of single reaction will be marked based on the result of union of this reaction in the child nodes. For example, if one reaction that both exists in two nodes x and y (eg. 1), the existence of their parent should also be 1 by computing the union result of the reaction of the two children; if one reaction that exist only in one child, the existence of this reaction in their parent cannot be decided and marked with uncertainty (eg. $\{0, 1\}$); if one reaction does not exist in any of the child nodes, this reaction should not exist in their parent (eg. 0). In the second process, the phylogenetic tree was traversed from top to the bottom and a gapless metabolic network was created at each node by filling the gaps remaining after the bottom-top Fitch pass. Reactions that were incorporated to fill the gaps should satisfy the conditions that the total cost of mutations was minimized (eg. if one reaction that disappears in the parent node but exists in the child node, a mutation has occured and the total cost would plus one). The main advantage of prediction method was that the whole procedure was constructed automatically. Mutation cost is produced when adding reactions that originally do not exist in the living species in order to construct the gapless metabolic network. Moreover, many plausible pathways are excluded by constructing maximum parsimony tree and the minimum mutation cost is calculated based on the tree structure. For example, when both the parent node and the child node don't exist the reaction, incorporating this reaction to the child node will produce one mutation (0 to 1 mutation) thus increasing the total mutation cost. Only the pathway with minimum mutation cost would be left thus resolved the commonly existing problem in computational-based method for predicting too

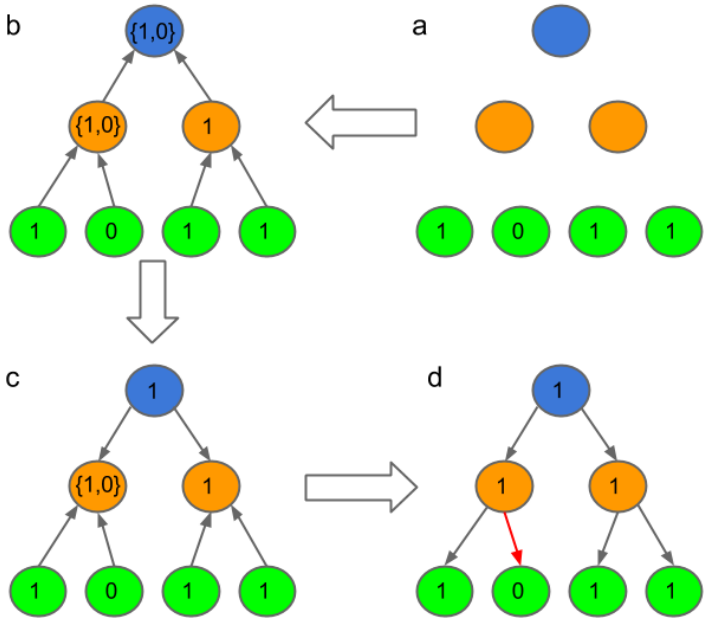


Figure 9: Fitch algorithm schematic diagram. (a) the existence of each reaction for a single species was first assigned to the leaf nodes (green); $(b), (c)$ Based on maximum parsimony rules (equation 13), the tree first construct from the bottom to the top; If the root node (blue) is not decided and marked with the set of $0, 1$, the value of this root node should be assigned with 1; (d) For internal nodes (yellow), if it is uncertainty and marked with the set of $0, 1$, the existence of this reaction for single internal node should be assigned with the same value of its parent node; Finally, the tree is reconstructed with one mutation (red arrow) in (d), which is the maximum parsimony tree.

many plausible pathways (Figure 10). However, the predicted pathway with mini-

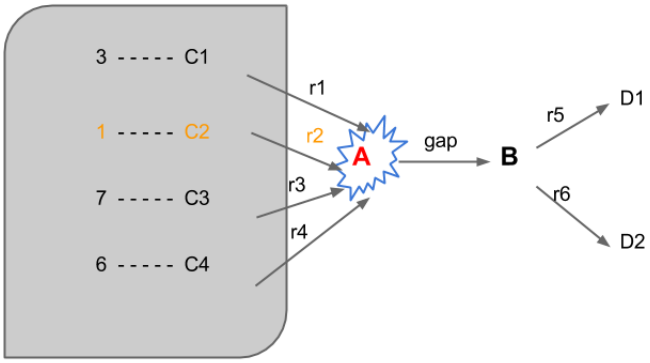


Figure 10: The gap filling process with minimum mutation cost. The example of metabolite A (red) cannot be reached in the model of one species based only on the Fitch algorithm from the parsimony tree. Dynamic programming is used and 4 reactions have been found to produce this metabolite with different mutation cost (eg. incorporate reaction $r1$ will increase the total mutation cost with 3). The reaction with minimum mutation cost (yellow) will be added.

mum mutation cost did not necessary mean this pathway was the real one existing in the species without carefully manual curation. The reasons were partially because the annotation error that recorded in KEGG and partially because the metabolism difference among species that some reactions should not incorporate in the species although these reactions were included in the pathway with minimum mutation cost.

3.2.2.2 The CoReCo algorithm

With more genomes and proteomes are available, the gap between sequence data and the metabolic model of species is increasing. Moreover, in my knowledge, there is not a tool can efficiently construct metabolic models for both living species and their ancestors. The CoReCo tool satisfies the goal and can predict metabolic models for different species at the same time, which offer a way for evolutionary study. There are three main steps (Figure 11): (1) sequence analysis, (2) construct probabilistic models by Pearl's poly tree algorithm [Pea88] based on the maximum parsimony tree and (3) create gapless metabolic models by atom mapping algorithm [HLMR11].

Sequence Analysis The first step of CoReCo method was processed by reciprocal alignments with BLAST (blastp 2.2.27+, evalue cutoff 10) that match 501619 protein sequences from 49 fungi against the Swiss-Prot database. E-value is used to evaluate the similarity between the query sequence and the target. Specifically, the smaller e-value between the query and the target, the bigger similarity it does.

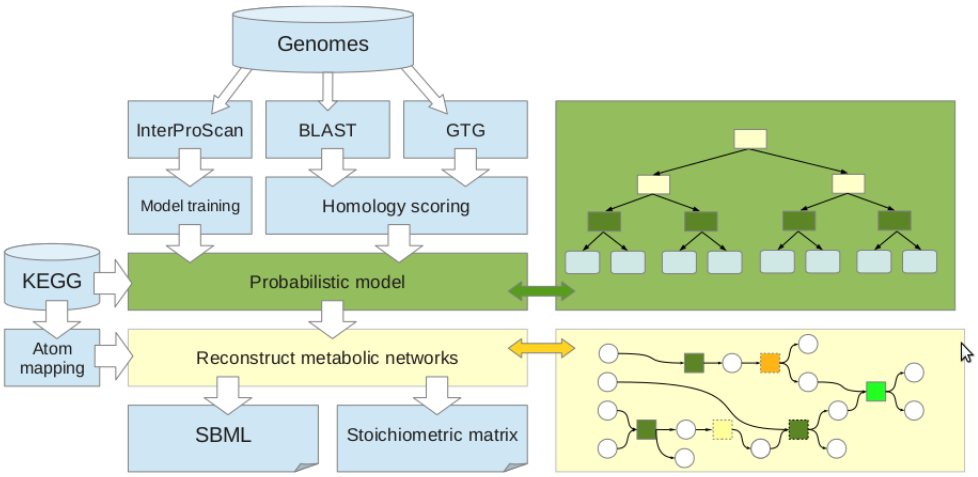


Figure 11: Overview of the reconstruction method by Pitkänen [PJH+14]. Sequence analysis: genomes of the target species is subjected to BLAST, GTG and InterProScan analyses to compute enzyme probabilities. Probabilistic model & Atom mapping: Plausible gapless metabolic networks are assembled based on integrated enzyme evidence. Finally reconstructed models are converted into SBML and generic stoichiometric matrix formats.

The result of BLAST alignment yielded the two e-value namely, $E(s,t)$ and $E'(s,t)$, where $E(s,t)$ is the forward Blast result by matching species sequences against the sequence from the Swiss-Prot database and $E'(s,t)$ is the reciprocal Blast result by matching sequences from the Swiss-Prot database against the sequence of the species. Next, two probabilities, $p(s,t)$ and $p'(s,t)$ respectively, were computed according to their e-value. The joint Blast score was designed for combining these two probabilities

$$B(s,t) = -\log\left(p(s,t) + p'(s,t) - p(s,t)p'(s,t)\right) \tag{14}$$

which described the similarity between sequences s and t. To detect the homology for distantly related species, GTG search was performed by extracting all of the conserved amino acid features (GTG feature) for each query sequence s. Then, the most similar sequence t from GTG database that belongs to the Swiss-Prot database and share the most GTG features with sequence s would be chosen. In particular, we cannot acquire annotations for the sequence not in Swiss-Prot database even if it shares most GTG features. In such case, it is good to choose the second or third similar sequence (eg. t) to express the query sequence s. GTG score was computed based on the shared GTG features between query sequence s and the best matches sequence t by

$$G(s,t) = \frac{|F(s) \cap F(t)|}{|F(s)|} \tag{15}$$

where, F(s) and F(t) are the number of GTG features in sequence s and t. Moreover, annotations (eg. EC number) would be added for each pair of $B(s,t)$ and $G(s,t)$ based on the Swiss-Prot ID of sequence t. Here is a sample result of GTG score (Table 6). Org is the species in short name; $SeqId$ is the sequence ID; $SeqGTGs$

Table 6: The sample result of GTG score.

Org	SeqId	SeqGTGs	MatchSeq	MatchNID	MatchGTGs	GTGScore	Ecs
Acla	A1C3S4	920	A1C3S4	1527257	920	1.000	?
Acla	A1C3S4	920	Q9ULF0	384665	679	0.738	?
Acla	A1C3S4	920	?	218106	651	0.708	?
Acla	A1C3S4	920	Q7VHV4	1575453	641	0.697	6.1.1.7
Acla	A1C3S4	920	Q7VQG3	1572842	640	0.696	6.1.1.7
Acla	A1C3S4	920	Q9RNN8	396147	636	0.691	6.1.1.7

is the number of GTG feature for the query sequence; $MatchSeq$ is the matched sequence in GTG database where the sequence will mark with "?" if it doesn't exist in the Swiss-Port database; $MatchNID$ is the GTG sequence ID; $MatchGTGs$ is the number of GTG features for the matched sequence; $GTGScore$ is the frequency between the number of query GTG against the number of matched GTG; ECs is the EC number annotated in Swiss-Prot database and record as "?" for unknown EC. In this case, the best match for the query sequence (A1C3S4) is Q7VHV4 with GTG score 0.697 and EC number 6.1.1.7. Finally, the best joint blast score and GTG score for each enzyme in species x would be selected by

$$B(e,x) = \max_{s \in Q, t \in T} B(s,t) \tag{16}$$

$$G(e,x) = \max_{s \in Q, t \in T} G(s,t) \tag{17}$$

where e is the enzyme in species x, Q is the proteome in species x and T is all of the protein sequence in Swiss-Prot database.

Reaction stoichiometry and Atom Mappings Reactions from KEGG have to be pretreatment in three steps before used in metabolic network reconstruction. The first step is to remove the reactions that are recorded as *general reaction* in KEGG. General reactions are specific group of reactions that do not have specific functions in metabolic pathways such as "$Dinucleotide + H_2O = 2Mononucleotide$" (KEGG number R00056). 191 general reactions have been filtered and 8664 reactions were left as the reaction pool for following reconstructions. In the second step, a balance formula was created for each of the reaction and the atom mapping algorithm by Heinonen *et al* [HLMR11] was used to compute the atom graph. Atom mapping can provide the best solution that match atoms and bonds from the reactants to the products based on the minimum edge edit distance [9]. The algorithm can be

[9]Given a pair of graphs G_1 and G_2, The edge edit distance $d_{EE}(G_1, G_2)$ is the minimum number of edge edit operations that is required to transform G_1 to G_2

divided into two processes. First, an optimal atom mapping was searched for. If the atom mapper failed to discover the optimal solution in a reasonable time, a heuristic approach was used to detect the non-optimal mapping.

Reconstruction steps After the preliminary steps of sequence processing, the metabolic network reconstruction of S.cerevisiae and P.pastoris were processed into two steps: CoReCo Phase I and CoReCo Phase II. In CoReCo Phase I, a new score (reco-phase1) was created for each EC number based on the BLAST and GTG score by Pearl polytree algorithm. In CoReCo Phase II, atom mapping algorithm was used to fill the reaction gaps and predict the metabolic networks of S.cerevisiae and P.pastoris.

CoReCo Phase I: Probabilistic Model Enzyme exists in living species (leaf nodes) was expressed in 1 and 0 if it does not exist. A maximum parsimony tree of each enzyme of the 49 fungi was constructed with 500 times based on the structure of the phylogenetic tree. By summing all mutations in 500 times for each edge from the ancestor to the child, the mutation probabilities were computed for each enzyme. The second step is to compute the conditional probability of each enzyme in the living species based on their GTG score and the joint Blast score by the kernel method (R package, Gaussian kernel method).

$$P(B|X) = \frac{1}{n} \sum_{x} \sum_{e \in R(x)} K(B - B(e, x)) \tag{18}$$

$$P(G|X) = \frac{1}{n} \sum_{x} \sum_{e \in R(x)} K(G - G(e, x)) \tag{19}$$

where n is the total number of summed score, $R(x)$ is the enzyme in the existing species, $B(e, x)$ and $G(e, x)$ are the Blast and GTG score for the enzyme e in species x, K is the kernel function summing to one. The EC number from the InterProScan (version 40.0) served as the golden standard in the kernel method. InterProScan [ZA01] combines 14 different Hidden Markov methods to predict the existence of each enzyme from 24117 protein signatures. The total number of 572 unique EC was predicted from the InterProScan process. Four conditional probabilities need to be trained by $P(B|X = 1)$, $P(B|X = 0)$, $P(G|X = 1)$ and $P(G|X = 0)$ where the probability of the enzyme existed in the predictions of both BLAST and InterProScan result was expressed as $P(B|X = 1)$; the probability of enzyme that predicted in BLAST but not in InterProScan result was exhibited as $P(B|X = 0)$; the probability of the enzyme existed in both GTG and InterProScan results was shown as $P(G|X = 1)$; the probability of the enzyme predicted only in GTG result was described as $P(G|X = 0)$. The distribution of the four estimated conditional probabilities was shown in Figure 12. Finally, the Bayesian network was constructed and the posterior probabilities $P(X|B, G)$ for each enzyme was calculated by *Pearl*

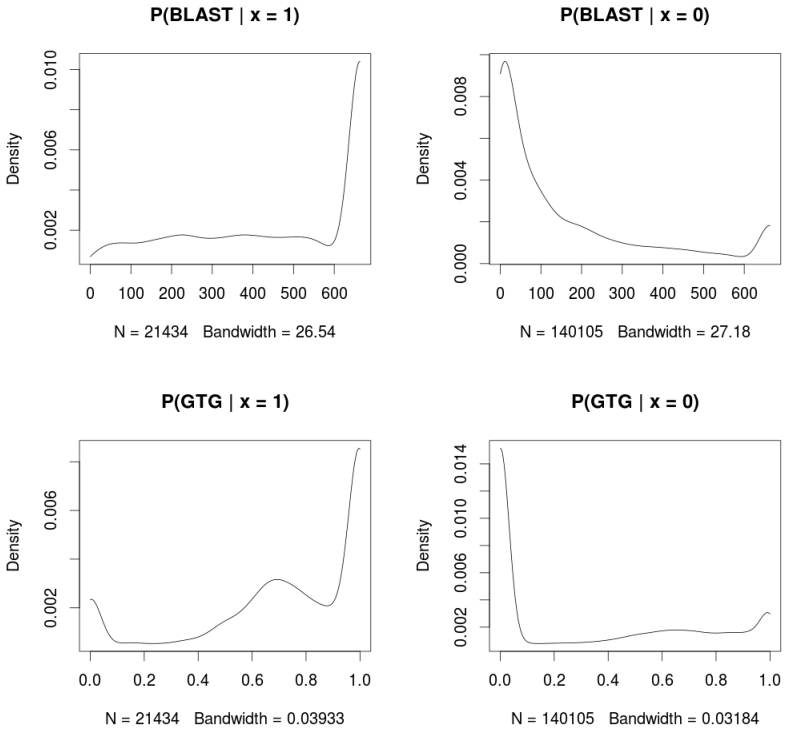


Figure 12: The conditional probability of the existing species. The range of the BLAST score (top left and top right) and the GTG score (bottom left and bottom right) are from 0 to 663.38 and from 0 to 1 respectively. The value on y axis is the density estimated by Gaussian kernel. N is the number of observations from BLAST or GTG results. *Bandwidth* is the estimated standard deviation based on the observations by Silverman's method (nrd0).

polytree algorithm [Pea88]. The Bayesian metabolic network was built for each enzyme based on the structure of the phylogenetic tree of multi-species (Figure 13) Pearl polytree algorithm uses different features (BLAST and GTG scores in our experiments) as input and computed the posterior probability of each ancestor from the priori probability of their children. These probabilities were served as input in the next phase where gapless metabolic networks were constructed for each fungi.

CoReCo Phase II: Gapless Atom-level Reconstruction Algorithm The atom-level reconstruction algorithm was based on the research [PJH+14]. The first step of the algorithm transferred the posterior probability $P(e, x)$ to a logarithmic cost for each reaction r by

$$C(r) = \begin{cases} -\log\left(P(e,x)\right) + \epsilon & \text{if } P(e,x) > \gamma \\ -\log\gamma + \epsilon & \text{others} \end{cases} \tag{20}$$

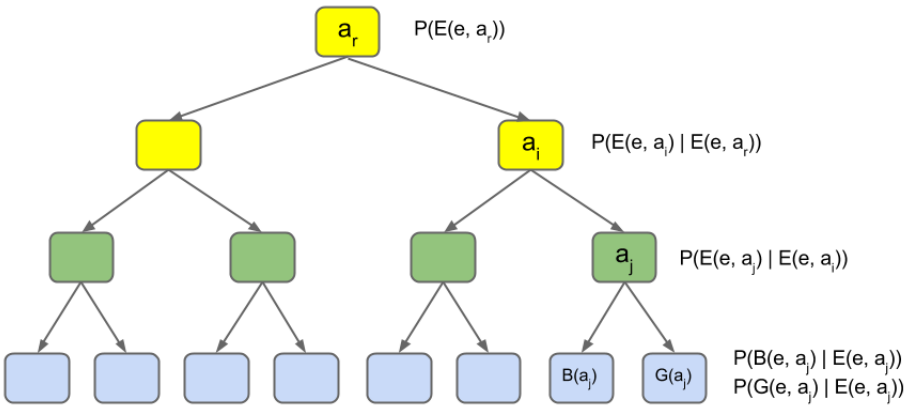


Figure 13: The Bayesian network of each enzyme by Pitkänen [PJH+14]. Prediction of existence of each enzyme is based on the phylogenetic tree of the multi-species. Three different nodes contain in the tree: the ancestral node (yellow), the existing species (green) and the Blast and GTG evidence nodes (blue).

where $P(e, x)$ is the posterior probability of enzyme e in species x annotated with reaction r, γ and ϵ are the minimum posterior probability of single reaction and the base cost that should be determined before test (eg. $\gamma = 1 \times 10^{-6}$ and $\epsilon = 1$ in my test.). In the second step, a heuristic approach [PJH+14] was designed in computing the gapless metabolic network. Reactions that satisfy the acceptance were arranged with increasing order based on their estimated costs. Additional costs would be added in each iteration when identifying the shortest path from nutrition to the metabolites of reactions that already incorporate in the metabolic network and summed the costs for the existing reactions and the new additional reactions. Two parameters are vital in this process: the acceptance threshold (α) and the rejection threshold (β). Each reaction that was smaller than the acceptance would be chosen and the reconstructed result of each pathway would be reported if the total amount of cost was less than the rejection threshold (equation 21).

$$C(N) = \sum_{r \in N} C(r) \tag{21}$$

Some reactions, which were solidly supported by sequence data (high posterior probability) but no gapfilling pathway has been found, would be flagged and the option of whether to incorporate these reactions into the metabolic model was judged by the researcher. The connectivity would be guaranteed if we rejected these reactions but some biological functions may be lost for the species; whereas the predicted accuracy would be increasing but the connectivity could not be maintained.

4 Analysis of CoReCo

The metabolic reconstruction algorithm, Comparative ReConstruction (CoReCo) developed by Pitkänen *et al* [PJH+14], was used to reconstruct the metabolic network of S.cerevisiae and P.pastoris. The reconstructed network of S.cerevisiae was evaluated by the yeast consensus model and the reconstructed model of P.pastoris was evaluated based on two models, namely iLC915 and PpaMBEL1254. In my thesis, five experiments were designed to evaluate the prediction ability of the CoReCo algorithm under different conditions. The first experiment was to evaluate the reconstruction accuracy of the predicted reactions by using receiver operating characteristic (ROC) curves for both P.pastoris and S.cerevisiae. Next, reconstructed models of P.pastoris and S.cerevisiae were computed under pooly sequenced data and the accuracy was tested by the areas under the curve (AUC). In the third experiment, the construct accuracy of S.cerevisiae and P.pastoris was computed from different phylogenetic trees and poorly sequenced data. The last two experiments were designed to test the two major parameters: acceptance (α) and rejection (β) with and without the gap penalty.

4.1 Original data

The reconstruction process of the CoReCo algorithm was computed based on proteomes of the 49 fungi. Proteomes of the 49 fungi were acquired from a previous study [PAR13] (Table 7). Two species were used for the metabolic network reconstruction: S.cerevisiae and P.pastoris. Recently, P.pastoris has been re-sequenced by De Schutter *et al* [DSLT+09] and the new proteome of the species was published online. The genome of P.pastoris is first segmented into small units and sequenced by 454/Roche sequencing separately, which is one of the next generation sequencing technology that can sequence the whole genome within several days. The assembly of the sequenced segments is finished by using shortgun method that merges two random segments when the P-value between them is smaller than e^{-20} from the Blastn [10] result.

Metabolites and reactions were collected from the KEGG database (the 2012 version). Swiss-Prot database and Global Trace Graph (GTG) (see 3.1.1 Resources) database were used for homologous searching of protein sequence. Swiss-Prot database (version 31.10.2012; download from Uniprot database) contains manually annotated records with information extracted from literature, which is highly reliable. GTG database [HMWH07] includes protein sequences from various databases (eg. Swiss-Prot, TREMBEL, PIR, PDB, WORMPEP and ENSEMBL) and was used to search the homologous proteins especially for the distantly related species.

[10]The tool computes the similarity between the query nuclear sequence and the target nuclear sequence and predicts the best match based on the P-value. There are several reasons that influence the P-value: query sequence length, match sequence length, size of the target database and the gap penalty.

Table 7: 49 fungi namelist

Full name	short name
Aspergillus niger	Anig
Trichoderma reesei	Tree
Puccinia graminis	Pgra
Sclerotinia sclerotiorum	Sscl
Debaryomyces hansenii	Dhan
Aspergillus nidulans	Anid
Ashbya gossypii	Agos
Fusarium oxysporum	Foxy
Neosartorya fischeri	Nfis
Laccaria bicolor	Lbic
Candida glabrata	Cgla
Encephalitozoon cuniculi	Ecun
Lodderomyces elongisporus	Lelo
Candida lusitaniae	Clus
Cryptococcus neoformans	Cneo
Phaeosphaeria nodorum	Pnod
Coprinus cinereus	Ccin
Sporobolomyces roseus	Sros
Mycosphaerella graminicola	Mgra
Chaetomium globosum	Cglo
Pichia stipitis	Psti
Coccidioides immitis	Cimm
Botrytis cinerea	Bcin
Fusarium graminearum	Fgra
Neurospora crassa	Ncra
Histoplasma capsulatum	Hcap
Yarrowia lipolytica	Ylip
Batrachochytrium dendrobatidis	Bden
Kluyveromyces lactis	Klac
Postia placenta	Ppla
Candida albicans	Calb
Uncinocarpus reesii	Uree
Phanerochaete chrysosporium	Pchr
Rhizopus oryzae	Rory
Saccharomyces cerevisiae	Scer
Schizosaccharomyces japonicus	Sjap
Pichia guilliermondii	Pgui
Fusarium verticillioides	Fver
Aspergillus clavatus	Acla
Aspergillus fumigatus	Afum
Aspergillus terreus	Ater
Aspergillus oryzae	Aory
Phycomyces blakesleeanus	Pbla
Nectria haematococca	Nhae
Magnaporthe grisea	Mgri
Pichia pastoris	Ppas
Ustilago maydis	Umay
Candida tropicalis	Ctro
Schizosaccharomyces pombe	Spom
Candida guilliermondii	Cgui

4.2 Reconstruction accuracy of P.pastoris and S.cerevisiae by ROC curves

Three metabolic models were used for testing the reconstruction accuracy of CoReCo in Phase I. PpaMBEL1254 [SGK+10] and iLC915 [CSA+12], both the predicted metabolic models of P.pastoris, contain different number of ECs and reactions due to different methods that have been used to construct these models; Yeast consensus model (version yeast_6.06) was used for testing the reconstruction accuracy for the metabolic model of S.cerevisiae. The basic statistics of the size of two models are shown in Table 8. Based on the ECs predicted from CoReCo for each species and

Name	ReactionNum	ECNum	MetaboliteNum
iLC915	1423	603	899
PpaMBEL1254	1202	425	1147
Yeast consensus model	1029	603	1351

Table 8: Characteristics of iLC915, PpaMBEL1254 and Yeast consensus model. The number of reactions, ECs and Metabolites are the values in ReactionNum, ECNum and MetaboliteNum respectively.

the ECs from the three metabolic models, the ROC curves were created for each species (Figure 14). Five different score methods were used to plot the ROC curve for a single species. Reco-phase1 is the EC score from CoReCo phase I calculated by Pearl polytree algorithm, which considered both Blast and GTG score for each EC with phylogenetic context. Reco-phase2 was the score for the EC calculated from CoReCo Phase I to CoReCo Phase II. Each EC in reco-phase1 has to satisfy the acceptance threshold (α) and the total amount of cost in each pathway should not more than rejection parameter (β) after using gap filling algorithm (see Methods). BLAST and GTG were the Blast score and the GTG score that directly result from blastp and GTG search result. NaiveBayes score is computed by combining BLAST and GTG score in the following equation

$$NaiveBayes(e, x) = 1/2B(s, t) + 1/2G(s, t) \tag{22}$$

where e is the enzyme of species x, B and G are the Blast and GTG score. By computing ROC curves and their area under the curve (AUC) values, the ROC curve of P.pastoris computed from reco-phase1 score was the best estimated by the iLC915 model and the PpaMBEL1254 model. The ROC curve of P.pastoris computed from reco-phase2 (red curves) verified by PpaMBEL1254 model was less accurate than the ROC curve with the same score tested by iLC915 model.

There are several reasons for these results. Comparison between these models indicated the unique EC number in iLC915 and PpaMBEL1254 was 603 and 445

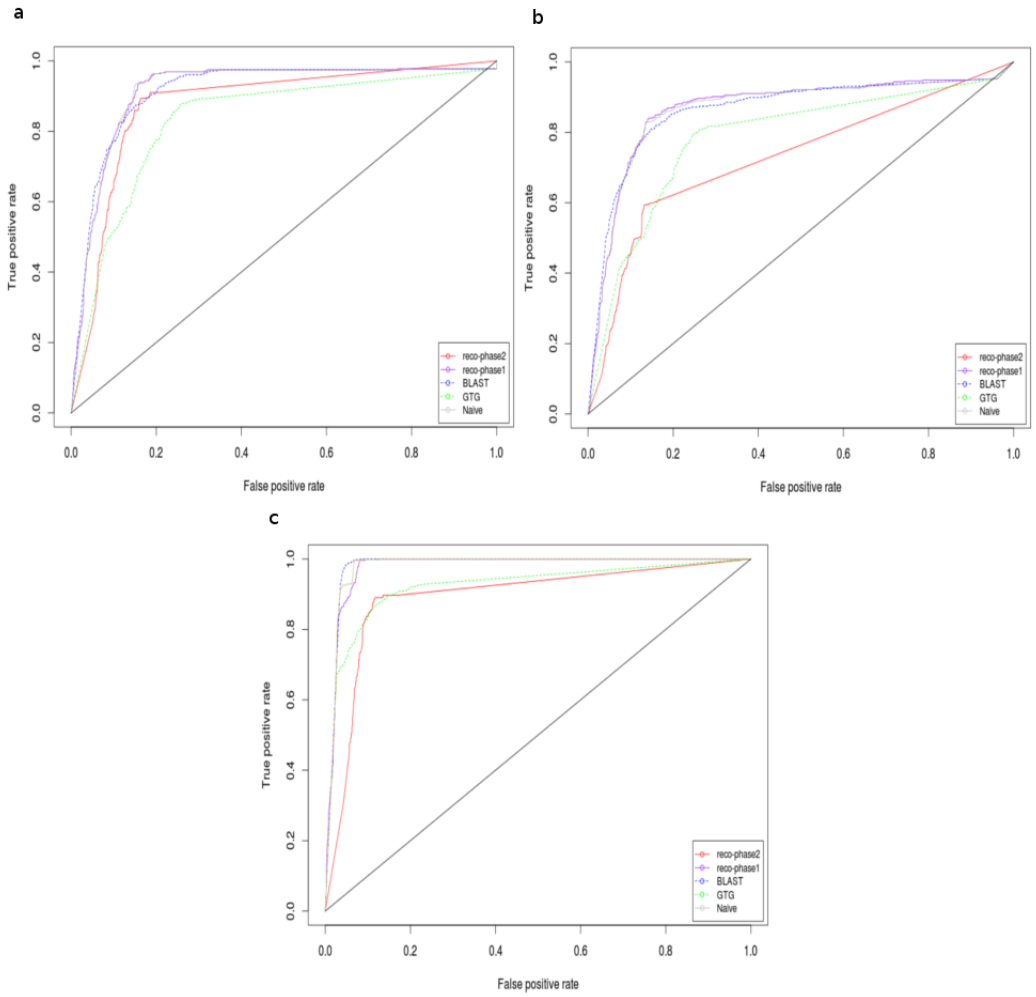


Figure 14: The ROC curves for P.pastoris and S.cerevisiae with different score methods. P.pastoris has been evaluated by iLC915 model (a) and the PpaMBEL1254 model (b). S.cerevisiae is tested in Yeast consensus model (c). x axis is the false positive rate and y is the true positive rate. There are 5 curves in each picture described by different score methods: reco-phase2 (red), reco-phase1 (purple), BLAST score(blue), GTG score (blue) and NaiveBayes (grey).

	iLC915	PpaMBEL1254	Share
ECNum	603	445	371

Table 9: Summary of unique and shared EC in iLC915 and PpaMBEL1254 models. Values in iLC915 and PpaMBEL1254 are the number of unique EC; share is the ECs that exist in both iLC915 and PpaMBEL1254 models.

respectively and the shared EC number in both two models was only 371 (Table 9). PpaMBEL1254 was constructed directly based on the homologous protein searching against the proteins in KEGG and TransportDB, which mean many EC numbers chosen from other species in PpaMBEL1254 may not be correct. Furthermore, parameters of homologous searching towards multi-species should be quite strict in order to discover the real ECs (eg. e-value in blastp searching), which means many ECs that should be included in the metabolic model are excluded in PpaMBEL1254. The iLC915 model was constructed based on the annotations of the metabolic model of S.cerevisiae (iIN800 [NJM$^+$08]). iIN800 was well annotated (included more reactions and metabolites) and the close relationship between S.cerevisiae and P.pastoris makes sure that homologous searching of the proteins in P.pastoris against the proteins in S.cerevisiae (iIN800) will produce more functional results (meaning more ECs are included in iLC915 and the ECs in iLC915 are real). Moreover, iLC915 incorporated more reactions that related in Methanol metabolism pathway, which is not included in PpaMBEL1254 model.

All ROC curves of S.cerevieiae presented in Figure 14 *c* were similar except the curve computed by reco-phase2 score (red). This is because S.cerevisiae is well annotated in both Swiss-prot and KEGG database so that using only BLAST or GTG searching against Swiss-prot and GTG database is enough to produce a good result. The ROC curve computed from reco-phase2 was less accurate than the other 4 curves in each figure. This is because some ECs with high posterior probability (low cost) chosen from the atom mapping algorithm in CoReCo Phase II were rejected to maintain the connectivity of the predicted metabolic network (eg. 129 reactions were necessary in order to fill the reaction gap R01108, which exceeded the reject threshold β although the posterior probability of R01108 was 0.87). Some ECs with low posterior probability (high cost) were included in order to fill reaction gaps in the network. Finally, the contingency table of S.cerevisiae and P.pastoris was created and verified by different metabolic models (Table 10). The sensitivity of P.pastoris by the iLC915 and PpaMBEL1245 models is 0.96 and 0.83. The specificity of P.pastoris by the iLC915 model and PpaMBEL1245 models is 0.84 and 0.85. The sensitivity and specificity of S.cerevisiae by the yeast consensus model are 0.97 and 0.92. In conclusion, the results of S.cerevisiae tested by ROC curves and contingency tables are better than P.pastoris.

Two main reasons cause the result. More information is available for S.cerevisiae than P.pastoris in both the Swiss-Prot database and the GTG database make the

Name	MetModel	TP	TN	FP	FN
P.pastoris	iLC915	579	2637	493	24
P.pastoris	PpaMBEL1245	373	2602	452	72
S.cerevisiae	yeast consensus model	586	2658	217	16

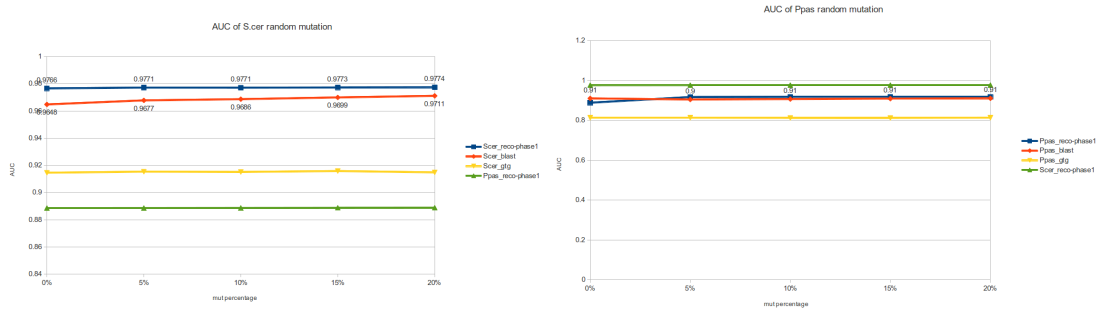
Table 10: Contingency table of P.pastoris and S.cerevisiae.$MetModel$ is the metabolic model for evaluation of the prediction result. TP is the number of ECs exist in both predicted model and tested model. TN is the number of ECs does not exist in both models. FP is the number of ECs that exists only in predicted model and FN is the number of EC predicted only in tested model.

BLAST and GTG search result of S.cerevisiae better than the result of P.pastoris. More annotations (eg. EC numbers) in yeast consensus model than the iLC915 and PpaMBEL1245 models also contribute to the better result of S.cerevisiae than P.pastoris.

4.3 Reconstruction accuracy with random sequence mutation and deletion

To test the performance of the CoReCo algorithm towards poorly sequenced data, the original protein sequence was processed by different percentage of random mutation and deletion. The original sequence was randomly mutated for S.cerevisiae and P.pastoris consisted from 5% to 20% of the total number of amino acids with 5% as one interval. Similarly, random deletion was made for the fungi deleted from 10% to 50% of the total number of genes with 10% as one interval. The CoReCo algorithm was reprocessed under these poorly sequenced data and the ROC curve computed from reco-phase1 score were used to evaluate the accuracy of S.cerevisiae and P.pastoris.

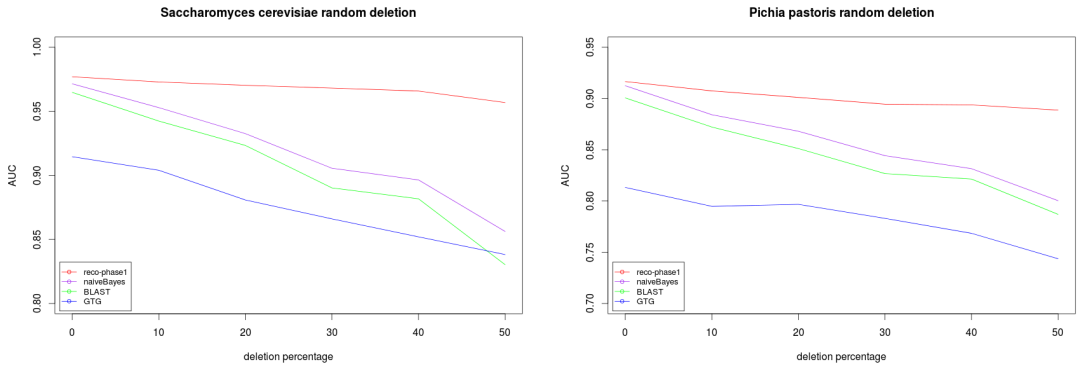
The results constructed under randomly mutated sequence of both S.cerevisiae and P.pastoris expressed the good performance of the CoReCo algorithm with little effects on random mutation (Figure 15). The AUCs under 20% of random mutation for both S.cerevisiae and P.pastoris were 0.97 and 0.91 respectively. The CoReCo algorithm was also stable towards random deletion of genes for S.cerevisiae and P.pastoris (Figure 16). The AUCs of S.cerevisiae and P.pastoris were verified under reference models by different percentage of random deletion from 10% to 50%. In both cases, the AUCs computed from reco-phase1 score for S.cerevisiae and P.pastoris were shown the best performance in each deletion rate. The AUCs of S.cerevisiae and P.pastoris with NaiveBayes score, BLAST score and GTG score were decreasing a lot under 50% deletion rate. However, the AUCs computed by reco-phase1 score in both S.cerevisiae and P.pastoris under 50% random deletion were almost the same as the AUCs from the original data.



(a) AUC of S.cerevisiae under different percentage of random mutation

(b) AUC of P.pastoris under different percentage of random mutation

Figure 15: The AUCs of S.cerevisiae and P.pastoris under different percentage of random mutation. x axis is the mutation percentage. y axis is the AUC range from 0 to 1. (a) describes the AUC of S.cerevisiae under different percentage of random mutations with reco-phase1 score (blue), Blast score (red) and GTG score (yellow). (b) is similar with (a) but describes the AUC of P.pastoris. In both pictures, three curves of one fungi are compared with the AUC computed from reco-phase1 score (green) of the other fungi.



(a) S.cerevisiae

(b) P.pastoris

Figure 16: The AUCs of S.cerevisiae and P.pastoris under different percentage of random deletion. x axis is the deletion percentage. y axis is the AUC range from 0 to 1. (a) describes the AUC of S.cerevisiae with reco-phase1 score (red), Naive Bayes score (purple), Blast score (green) and GTG score (blue). (b) is similar with (a) but describes the AUC of P.pastoris.

4.4 Reconstruction accuracy with varying sizes of the phylogenetic tree

The posterior probability of each enzyme, computed by Pearl polytree algorithm, is decided by the ancestor (parent node) and two children (child node) in the phylogenetic tree. The phylogenetic tree of the 49 fungi was based on a previous study [PAR13](Figure 17). Seven phylogenetic trees of S.cerevisiae (Figure 18) with dif-

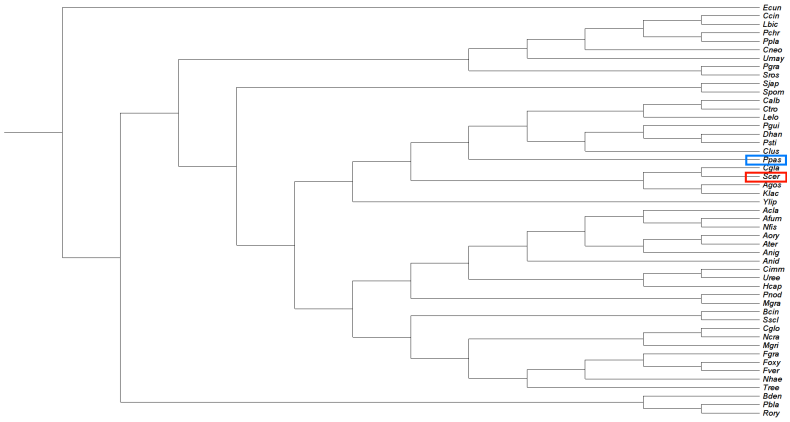


Figure 17: The phylogenetic tree of 49 fungi. The phylogenetic tree is constructed based on the evolutionary relationship between fungi. The name of each fungi is described by the short name (additional files). Specifically, S.cerevisiae (in red rectangle) and P.pasotirs (in blue rectangle) are named as ***Scer*** and ***Ppas***.

ferent number of neighbors were created based on the evolutionary relationship described in Figure 17. To evaluate the performance of the CoReCo algorithm under different neighbors of phylogenetic tree towards poorly sequenced data, metabolic networks of S.cerevisiae were constructed with each phylogenetic tree and each deletion rate. Similar process was also processed for P.pastoris. In the end, the AUC was calculated under specific conditions (different trees and different deletion rates) for both S.cerevisiae and P.pastoris and described by response surface (Figure 19). The predicted AUCs in both S.cerevisiae and P.pastoris were decreasing when the percentage of random deletion was increasing. Moreover, the CoReCo algorithm could predict well in the extreme cases (eg. the predicted AUCs of S.cerevisiae and P.pastoris were more than 0.95 at 50% random deletion). The best AUC of S.cerevisiae in each deletion rate was shown under the phylogenetic tree with 12 neighbors (red line). This is because the 12 fungi are more close-related with S.cerevisiae (see Figure 17), which means similar functions (eg. genes and reactions) in these fungi should also exist in S.cerevisiae and the prediction based on the phylogenetic tree was more accurate. The AUC was not increasing but became more stable when adding more fungi into the metabolic reconstruction process. Similar results were also observed in the response surface of P.pastoris. The maximum AUC in each deletion rate was presented in the phylogenetic tree with four neighbors that are more close-related with P.pastoris.

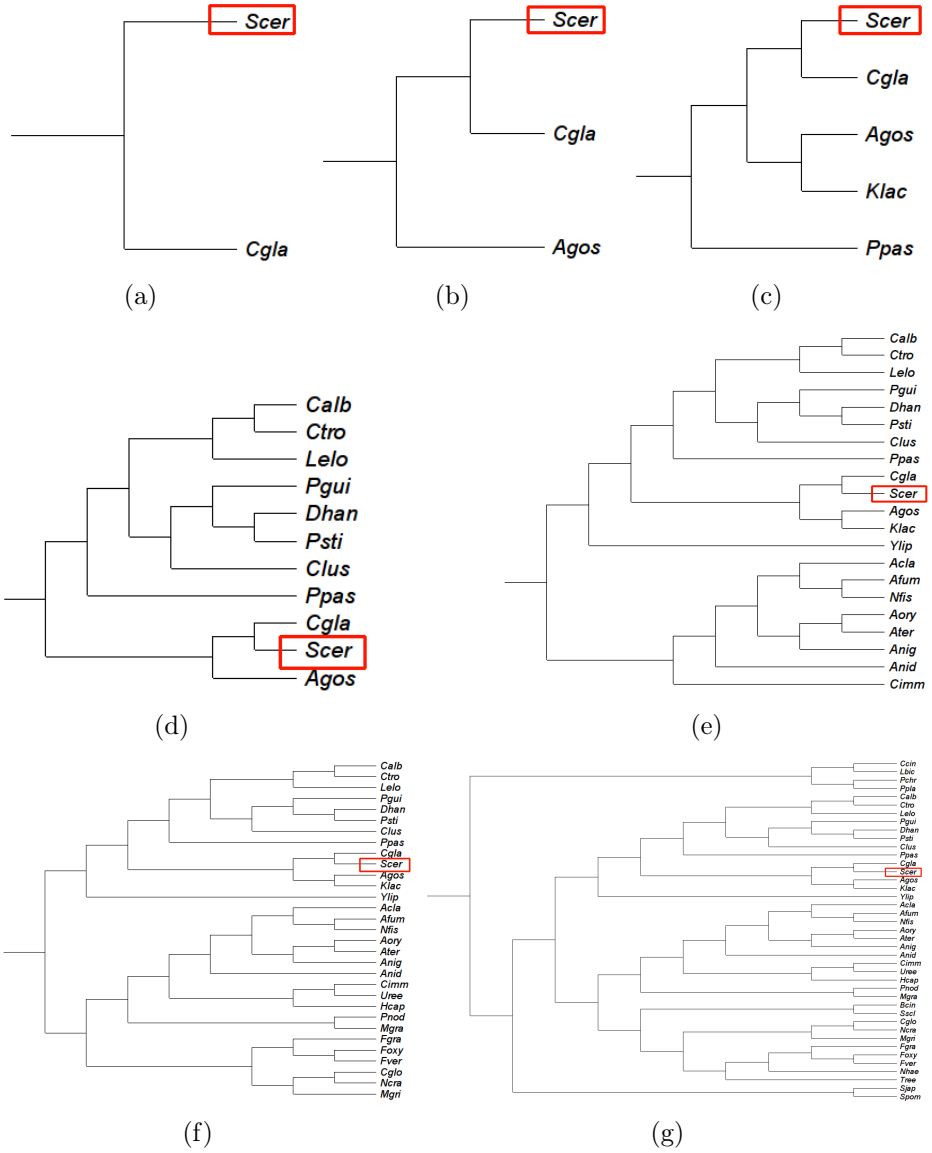
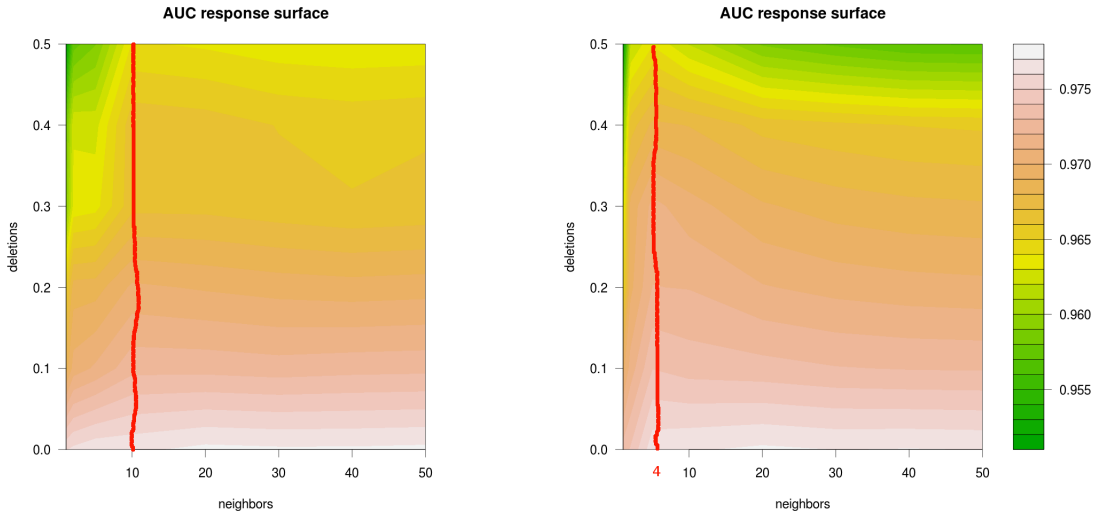


Figure 18: Phylogenetic trees of S.cerevisiae with different number of neighbors. Seven phylogenetic trees are created with 1 neighbor (a), 2 neighbors (b), 5 neighbors (c), 10 neighbors (d), 20 neighbors (e), 30 neighbors (f) and 40 neighbors (g).



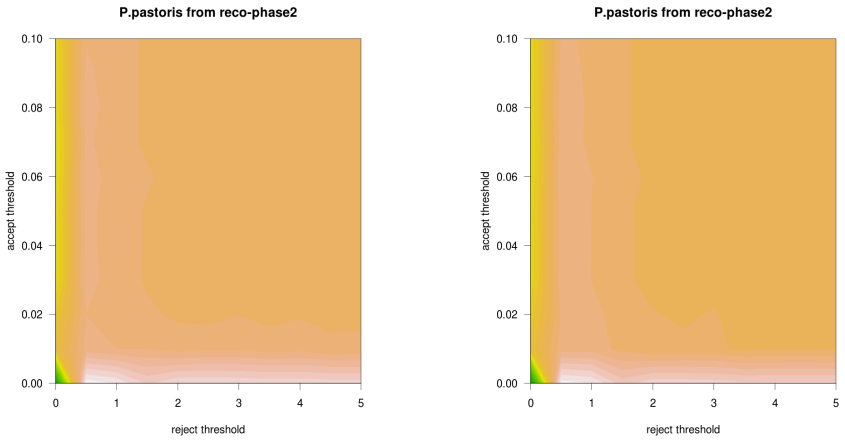
(a) S.cerevisiae (b) P.pastoris

Figure 19: Response surface of S.cerevisiae and P.pastoris. x axis is the number of neighbors of the phylogenetic tree. y axis is the percentage of random deletion. Metabolic networks are constructed under each conditions and the AUCs are computed for both S.cerevisiae (a) and P.pastoris (b).

4.5 Reconstruction results with different acceptance and rejection parameters

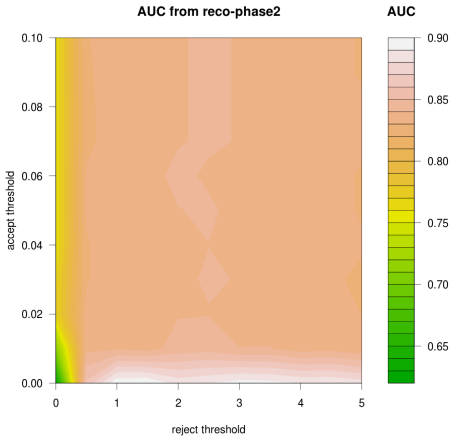
To evaluate the influence of the acceptance and rejection parameters in the CoReCo algorithm, the metabolic network of S.cerevisiae and P.pastoris was constructed under different combinations of acceptance and rejection parameters. Two experiments were completed. In the first test, all reactions predicted from CoReCo Phase II were assumed in the predicted metabolic networks without considering the connectivity. In the second test, only the gapless reactions were considered into the predicted metabolic networks and new response surface was created.

Evaluation of the reconstruction results without considering connectivity The first test was completed by incorporating all reactions predicted from CoReCo Phase II as input to compute the AUC under different acceptance and rejection (Figure 20). Acceptance (α) was chosen from 0 to 0.1 with 0.01 as one interval. Rejection (β) was ranging from 0 to 5 with 0.5 as one interval. The metabolic networks of both S.cerevisiae and P.pastoris were constructed under each acceptance and rejection. The AUCs of P.pastoris tested for both the iLC915 and PpaMBEL1245 model were similar under different acceptance and rejection. The maximum AUC of P.pastoris was 0.9 predicted with parameters $\alpha = 0$ and $\beta = 1$. This is because with zero acceptance, only the reactions with zero cost (or posterior probability equals one) were selected and the majority of the reactions with cost more than zero were rejected. Most Reaction threshold with zero cost should be the real reactions



(a) P.pastoris tested by iLC915

(b) P.pastoris tested by PpaMBEL1245



(c) S.cerevisiae

Figure 20: Response surface of S.cerevisiae and P.pastoris under different acceptance and rejection. x axis is the value of rejection. y axis is the value of accept. Metabolic networks are constructed under each conditions and the AUCs are computed for both P.pastoris (a) and (b) and S.cerevisiae (c) tested by different models.

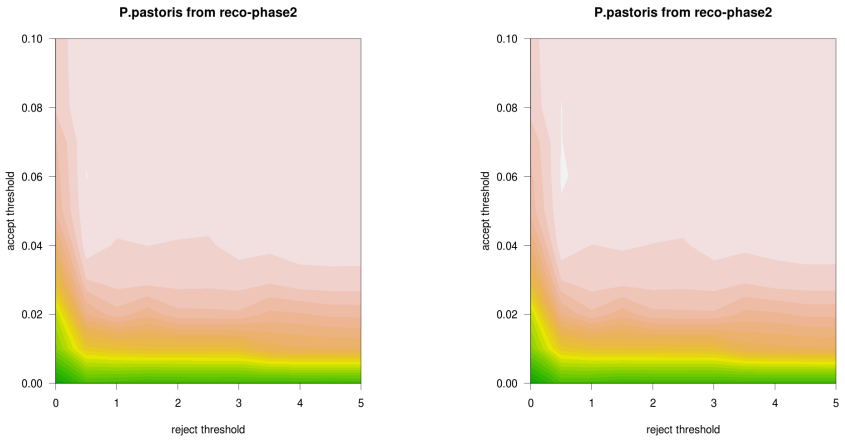
that existed in both the iLC915 and PpaMBEL1245 model so that the best AUC occurred. The minimum AUC of P.pastoris was 0.61 with paramters $\alpha = 0$ and $\beta = 0$. Similar results were also produced for S.cerevisiae. However, many reactions included in the predicted model had gaps. The predicted model, although with high accuracy when acceptance equals to zero, can not be directly utilized in engineering projects due to the reaction gaps. It was reasonable increasing the acceptance to create a gapless metabolic network with less accuracy.

Evaluation of the reconstruct results with gap penalty The second test was to test the accuracy of the predicted model for both S.cerevisiae and P.pastoris with gap penalty. Metabolic networks were also constructed based on the different acceptance and rejection parameter. Reactions from CoReCo Phase II were considered into the predicted model only if the reactions were flagged with gapless. The AUC of each predicted network was computed and depicted in response surface (Figure 21). The minimum AUC of P.pastoris and S.cerevisiae was shown with parameters $\alpha = 0$ and $\beta = 0$, which was different from the first test. This is because in the first test, the majority of reactions with more than zero cost were excluded and the atom mapping algorithm could not construct the gapless metabolic networks by only considering the reactions with zero cost. The maximum AUC of S.cerevisiae was 0.88 with parameters $\alpha = 0.05$ and $\beta = 2$. The maximum AUC of P.pastoris tested by the iLC915 and PpaMEBL1245 model was 0.84 and 0.86 with the parameters $\alpha = 0.06$ and $\beta = 0.05$ respectively. The CoReCo algorithm was proven to have the ability to predict well under a wide range of accept and rejection. For example, for any acceptance and rejection that are bigger than 0.04 and 0.5 respectively, the AUCs of both S.cerevisiae and P.pastoris were larger than 0.84.

5 Discussion

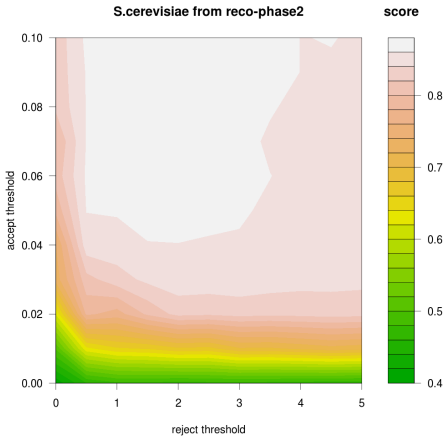
The CoReCo algorithm can predict the metabolic networks from poorly sequenced data with high accuracy. S.cerevisiae and P.pastoris were used here for metabolic network reconstruction. The ROC computed by the reference models, Yeast consensus model for S.cerevisiae, and two new models, iLC915 and PpaMBEL1245 models for P.pastoris, shown good results. Moreover, the AUCs were computed under different percentage of random mutation and deletion for both S.cerevisiae and P.pastoris with high accuracy means CoReCo performs well even for poorly sequenced data.

The CoReCo is an efficient tool to study the evolutionary relationship of unknown species and predicts the metabolic network of the hypothetical ancestor. The whole pipeline of model reconstruction of one fungi would be finished within 24 hours. Reconstruction process of the CoReCo method was built on phylogeny and the evidence of each enzyme was extended by computing the posterior probability of each EC in both the livings and ancestor species. The method offers a new orientation in the comparative genomics that evolution can be studied by comparison between



(a) P.pastoris tested by iLC915

(b) P.pastoris tested by PpaMBEL1245



(c) S.cerevisiae

Figure 21: Response surface of S.cerevisiae and P.pastoris under differnet acceptance and rejection. x axis is the value of rejection. y axis is the value of accept. Metabolic networks are constructed under each conditions and the AUCs are computed for both P.pastoris (a) and (b) and S.cerevisiae (c) tested by different models.

metabolic networks rather than compared under the sequence level (eg. genome comparison or proteome comparison).

The most time consuming step in the CoReCo algorithm is the InterProScan process and the reconstruction accuracy is severely influenced by the result from InterProScan. The EC numbers predicted from InterProScan were significantly less than the EC numbers in the test model. The number of EC predicted from InterProScan was 336 for P.pastoris while the EC number in the iLC915 and PpaMBEL1245 model were 603 and 445; the number of EC predicted from InterProScan was 347 for S.cerevisiae while the EC number in the yeast consensus model was 603. The difference of EC number between the results from InterProScan and the reference model (eg. iLC915) influence the reconstruction accuracy. Further test was made by re-computing the probability density using reference model as training sets. The result constructed by the smaller set from the InterProScan process would be different compared to the probability density trained with the reference model. It was more significant when the EC in the InterProScan was severely different to the EC in the test model described by the BLAST score or the GTG score. This is suggest using reference model as training when reconstructing the model could give a better construction result. We can use reference model of each species instead of the results from InterProScan as training sets to reconstruct the model when more reference models are available. This will dramatically decrease the time of model reconstruction.

Curation is still needed to produce a better reconstruction network. Reactions marked with gaps in the final results need to be decided whether to incorporate into the prediction. For example, for the reaction marked with "gaps" and small cost, discard the reaction may produce functional losses while incorporating the reaction will bring gaps into the model. Experiments can be developed for testing the reactions marked with gaps and small cost. We can decide whether to incorporate these reactions by extensive literature searching. In the end, other methods can be used to simulate biomass under the predicted metabolic network to see if the productivity is significantly influenced by the reaction gaps.

References

ABJ13 Arico, C., Bonnet, C. and Javaud, C., N-glycosylation humanization for production of therapeutic recombinant glycoproteins in saccharomyces cerevisiae. In *Glycosylation Engineering of Biopharmaceuticals*, Springer, 2013, pages 45–57.

AGM+90 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J., Basic local alignment search tool. *Journal of molecular biology*, 215,3(1990), pages 403–410.

ALS+13 Agren, R., Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I. and Nielsen, J., The raven toolbox and its use for generating a genome-scale metabolic model for penicillium chrysogenum. *PLoS computational biology*, 9,3(2013), page e1002980.

And92 Anderson, A., Yeast genome project:300,000 and counting. *Science*, 256,5056(1992), page 462.

ash

Bai00 Bairoch, A., The enzyme database in 2000. *Nucleic acids research*, 28,1(2000), pages 304–305.

BBA+03 Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I. et al., The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic acids research*, 31,1(2003), pages 365–370.

BFC+90 Beesley, K., Francis, M., Clarke, B., Beesley, J., Dopping-Hepenstal, P., Clare, J., Brown, F. and Romanos, M., Expression in yeast of amino-terminal peptide fusions to hepatitis b core antigen and their immunological properties. *Nature Biotechnology*, 8,7(1990), pages 644–649.

BHNSPP13 Bomholt, J., Hélix-Nielsen, C., Scharff-Poulsen, P. and Pedersen, P. A., Recombinant production of human aquaporin-1 to an exceptional high membrane density in saccharomyces cerevisiae. *PloS one*, 8,2(2013), page e56431.

BKN05 Borodina, I., Krabben, P. and Nielsen, J., Genome-scale analysis of streptomyces coelicolor a3 (2) metabolism. *Genome Research*, 15,6(2005), pages 820–829.

BLT93 Boguski, M. S., Lowe, T. M. and Tolstoshev, C. M., dbest-database for "expressed sequence tags". *Nature genetics*, 4,4(1993), pages 332–333.

CAD12 Caspi, R., Altman, T. and Dreher, L. A., The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Research*, 40,D1(2012), pages D742–D753.

CC00 Cereghino, J. L. and Cregg, J. M., Heterologous protein expression in the methylotrophic yeast pichia pastoris. *FEMS Microbiology Reviews*, 24,1(2000), pages 45–66.

CHP+13 Costa, C., Henriques, A. S., Pires, C., Nunes, J., Ohno, M., Chibana, H., SÃ¡-Correia, I. and Teixeira, M. C., The dual role of candida glabrata drug:h+ antiporter cgaqr1 (orf cagl0j09944g) in antifungal drug and acetic acid resistance. *Frontiers in Microbiology*, 4,170(2013).

CJF+13 Coutinho, L. C. d. A., Jesus, A. L. S. d., Fontes, K. F. L. d. P., Coimbra, E. C., Mariz, F. C., Freitas, A. C. d., Maia, R. d. C. C. and Castro, R. S. d., Production of equine infectious anemia virus (eiav) antigen in pichia pastoris. *Journal of virological methods*, 191,2(2013), pages 95–100.

CSA+12 Caspeta, L., Shoaie, S., Agren, R., Nookaew, I. and Nielsen, J., Genome-scale metabolic reconstructions of Pichia stipitis and Pichia pastoris and in silico evaluation of their potentials. *BMC Systems Biology*, 6,1(2012), page 24.

Dar12 Darvishi, F., Expression of native and mutant extracellular lipases fromyarrowia lipolytica in saccharomyces cerevisiae. *Microbial Biotechnology*, 5,5(2012), pages 634–641.

DH05 Daly, R. and Hearn, M. T. W., Expression of heterologous proteins in pichia pastoris: a useful experimental tool in protein engineering and production. *Journal of Molecular Recognition*, 18,2(2005), pages 119–138.

DÖHN08 David, H., Özçelik, İ. Ş., Hofmann, G. and Nielsen, J., Analysis of aspergillus nidulans metabolism at the genome-scale. *BMC genomics*, 9,1(2008), page 163.

DS96 Dosanjh, A. and Stone, K., Why pichia pastoris. *Retrieved June*, 15, page 2004.

DSLT+09 De Schutter, K., Lin, Y.-C., Tiels, P., Van Hecke, A., Glinka, S., Weber-Lehmann, J., Rouzé, P., Van de Peer, Y. and Callewaert, N., Genome sequence of the recombinant protein production host pichia pastoris. *Nature biotechnology*, 27,6(2009), pages 561–566.

FGCS13 Fang, W., Gao, H., Cao, Y. and Shan, A., Cloning and expression of a xylanase xynb from aspergillus nigeria-001 in pichia pastoris. *Journal of Basic Microbiology*. URL `http://dx.doi.org/10.1002/jobm.201300078`.

GT10 Gudmundsson, S. and Thiele, I., Computationally efficient flux variability analysis. *BMC bioinformatics*, 11,1(2010), page 489.

HDB+10 Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B. and Stevens, R. L., High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*, 28,9(2010), pages 977–982.

HH03 Heger, A. and Holm, L., Exhaustive enumeration of protein domain families. *Journal of molecular biology*, 328,3(2003), pages 749–767.

HHL+81 Hitzeman, R. A., Hagie, F. E., Levine, H. L., Goeddel, D. V., Ammerer, G. and Hall, B. D., Expression of a human gene for interferon in yeast. *Nature*, 293,5835(1981), pages 717–722.

HLMR11 Heinonen, M., Lappalainen, S., Mielikäinen, T. and Rousu, J., Computing atom mappings for biochemical reactions without subgraph isomorphism. *Journal of Computational Biology*, 18,1(2011), pages 43–58.

HMWH07 Heger, A., Mallick, S., Wilton, C. and Holm, L., The global trace graph, a novel paradigm for searching protein sequence databases. *Bioinformatics*, 23,18(2007), pages 2361–2367.

HXH+13 Hao, J., Xu, L., He, H., Du, X. and Jia, L., High-level expression of staphylococcal protein a in pichia pastoris and purification and characterization of the recombinant protein. *Protein Expression and Purification*, 90,2(2013), pages 178 – 185.

HZM12 Hong, S.-Y., Zurbriggen, A. and Melis, A., Isoprene hydrocarbons production upon heterologous transformation of saccharomyces cerevisiae. *Journal of applied microbiology*, 113,1(2012), pages 52–65.

ISI+13 Iram, N., Shah, M., Ismat, F., Habib, M., Iqbal, M., Hasnain, S. and Rahman, M., Heterologous expression, characterization and evaluation of the matrix protein from newcastle disease virus as a target for antiviral therapies. *Applied Microbiology and Biotechnology*, pages 1–11. URL `http://dx.doi.org/10.1007/s00253-013-5043-2`.

JBD+13 Jiménez, J. J., Borrero, J., Diep, D. B., Gútiez, L., Nes, I. F., Herranz, C., Cintas, L. M. and Hernández, P. E. *Journal of Industrial Microbiology & Biotechnology*, pages 1–17.

KDM07 Kumar, V. S., Dasika, M. S. and Maranas, C. D., Optimization based automated curation of metabolic reconstructions. *BMC bioinformatics*, 8,1(2007), page 212.

KG00 Kanehisa, M. and Goto, S., Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28,1(2000), pages 27–30.

KOMK+05 Karp, P. D., Ouzounis, C. A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., Tsoka, S., Darzentas, N., Kunin, V. and López-Bigas, N., Expansion of the biocyc collection of pathway/genome databases to 160 genomes. *Nucleic acids research*, 33,19(2005), pages 6083–6089.

KPR02 Karp, P. D., Paley, S. and Romero, P., The pathway tools software. *Bioinformatics*, 18,suppl 1(2002), pages S225–S232.

LMCK07 LEGRAS, J.-L., Merdinoglu, D., CORNUET, J. and Karst, F., Bread, beer and wine: Saccharomyces cerevisiae diversity reflects human history. *Molecular Ecology*, 16,10(2007), pages 2091–2102.

Met09 Metzker, M. L., Sequencing technologies-the next generation. *Nature Reviews Genetics*, 11,1(2009), pages 31–46.

MNdC+10 Montagud, A., Navarro, E., de Córdoba, P. F., Urchueguía, J. and Patil, K., Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium. *BMC systems biology*, 4,1(2010), page 156.

NJM+08 Nookaew, I., Jewett, M. C., Meechai, A., Thammarongtham, C., Laoteng, K., Cheevadhanarak, S., Nielsen, J. and Bhumiratana, S., The genome-scale metabolic model iin800 of saccharomyces cerevisiae and its validation: a scaffold to query lipid metabolism. *BMC Systems Biology*, 2,1(2008), page 71.

NVI+90 Nagadish, M. N., Vaughan, P. R., Irving, R. A., Azad, A. A. and Macreadie, I. G., Expression and characterization of infectious bursal disease virus polyprotein in yeast. *Gene*, 95,2(1990), pages 179–186.

OP12 Orth, J. D. and Palsson, B., Gap-filling analysis of the ijo1366 escherichia coli metabolic network reconstruction for discovery of metabolic functions. *BMC systems biology*, 6,1(2012), page 30.

Ost03 Osterman, A., Missing genes in metabolic pathways: a comparative genomics approach. *Current Opinion in Chemical Biology*, ,1, page 7.

PAR13 Pitkänen, E., Arvas, M. and Rousu, J., Reconstructing gapless ancestral metabolic networks. In *Biomedical Engineering Systems and Technologies*, Springer, 2013, pages 126–140.

Pea88 Pearl, J., *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

PJH+14 Pitkänen, E., Jouhten, P., Hou, J., Syed, M. F., Blomberg, P., Kludas, J., Oja, M., Holm, L., Penttilä, M., Rousu, J. and Arvas, M., Comparative genome-scale reconstruction of gapless metabolic networks for present and ancestral species. *PLoS Computational Biology*, 10,2(2014), page e1003465.

PN05 Patil, K. R. and Nielsen, J., Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proceedings of the National Academy of Sciences of the United States of America*, 102,8(2005), pages 2685–2689.

PPS+13 Pingitore, P., Pochini, L., Scalise, M., Galluccio, M., Hedfalk, K. and Indiveri, C., Large scale production of the active human ASCT2 (slc1a5) transporter in pichia pastoris - functional and kinetic asymmetry revealed in proteoliposomes. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1828,9(2013), pages 2238 – 2246.

PRA11 Pitkänen, E., Rousu, J. and Arvas, M., Minimum mutation algorithm for gapless metabolic network evolution. *BIOINFORMATICS*, 2011, pages 28–38.

PRU10 Pitkänen, E., Rousu, J. and Ukkonen, E., Computational methods for metabolic reconstruction. *Current opinion in biotechnology*, 21,1(2010), pages 70–77.

RBCTR04 Reverter-Branchat, G., Cabiscol, E., Tamarit, J. and Ros, J., Oxidative damage to specific proteins in replicative and chronological-aged saccharomyces cerevisiae common targets and prevention by calorie restriction. *Journal of Biological Chemistry*, 279,30(2004), pages 31983–31989.

RSC92 Romanos, M. A., Scorer, C. A. and Clare, J. J., Foreign gene expression in yeast: a review. *Yeast*, 8,6(1992), pages 423–488.

SGK+10 Sohn, S. B., Graf, A. B., Kim, T. Y., Gasser, B., Maurer, M., Ferrer, P., Mattanovich, D. and Lee, S. Y., Genome-scale metabolic model of methylotrophic yeast pichia pastoris and its use for in silico analysis of heterologous protein production. *Biotechnology journal*, 5,7(2010), pages 705–715.

SMG97 Sinclair, D. A., Mills, K. and Guarente, L., Accelerated aging and nucleolar fragmentation in yeast sgs1 mutants. *Science*, 277,5330(1997), pages 1313–1316.

SPLP93 Strand, M., Prolla, T. A., Liskay, R. M. and Petes, T. D., Destabilization of tracts of simple repetitive dna in yeast by mutations affecting dna mismatch repair. *Nature*, 365,6443(1993), pages 274–276.

SVC02 Segre, D., Vitkup, D. and Church, G. M., Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences*, 99,23(2002), pages 15112–15117.

TP10 Thiele, I. and Palsson, B., A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5,1(2010), pages 93–121.

vHS90 van Heeke, G. and Schuster, S. M., Expression of human asparagine synthetase in saccharomyces cerevisiae. *Protein engineering*, 3,8(1990), pages 739–744.

VOH+08 Vongsangnak, W., Olsen, P., Hansen, K., Krogsgaard, S. and Nielsen, J., Improved annotation through genome-scale metabolic modeling of aspergillus oryzae. *BMC genomics*, 9,1(2008), page 245.

WJLW13 Wang, M., Jiang, S., Liu, X. and Wang, Y., Expression, purification, and immunogenic characterization of epstein-barr virus recombinant ebna1 protein in pichia pastoris. *Applied Microbiology and Biotechnology*, 97,14(2013), pages 6251–6262.

WWL+13 Wang, F., Wu, M., Liu, W., Shen, Q., Sun, H. and Chen, S., Expression, purification, and lipolytic activity of recombinant human serum albumin fusion proteins with one domain of human growth hormone in pichia pastoris. *Biotechnology and Applied Biochemistry*, 60,4(2013), pages 405–411.

YCSZ13 Yang, X., Cong, H., Song, J. and Zhang, J., Heterologous expression of an aspartic protease gene from biocontrol fungus trichoderma asperellum in pichia pastoris. *World Journal of Microbiology and Biotechnology*, pages 1–8.

YLL+13 You, L.-F., Liu, Z.-M., Lin, J.-F., Guo, L.-Q., Huang, X.-L. and Yang, H.-X., Molecular cloning of a laccase gene from ganoderma lucidum and heterologous expression in pichia pastoris. *Journal of Basic Microbiology*. URL `http://dx.doi.org/10.1002/jobm.201200808`.

ZA01 Zdobnov, E. M. and Apweiler, R., Interproscan–an integration platform for the signature-recognition methods in interpro. *Bioinformatics*, 17,9(2001), pages 847–848.